ADAPTIVE BEHAVIOR OF IMPATIENT CUSTOMERS IN INVISIBLE QUEUES

RESEARCH THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN ELECTRICAL ENGINEERING

ESTER ZOHAR

Submitted to the senate of the Technion - Israel institute of technology

Shvat, 5761 HAIFA January 2001

This research was done under the supervision of Prof. Nahum Shimkin and Prof. Avishai Mandelbaum in the Faculty of Electrical Engineering.

I would like to thank Prof. Nahum Shimkin and Prof. Avishai Mandelbaum for the help and support they gave me. Without their help this work would not have been finished.

The generous financial help of the Technion is gratefully acknowledged

Contents

1 Abstract 1

2 Notations list 3

3 Introduction 4

3.1 Background and Motivation 4

3.2 Assumptions and Results 5

3.3 A learning model - formulation and simulation 6

3.4 Contents 7

4 Related research 8

4.1 Related work on queueing systems 8

4.2 Psychology related articles 9

5 Model Formulation 14

5.1 System model 14

5.2 The patience function 16

5.3 Rational abandonment 23

5.3.1 The utility function 23

5.3.2 The cost function 25

5.3.3 Properties of optimal decisions 28

6 Equilibrium Uniqueness 31

7 Survival Analysis 34

7.1 Censoring in waiting time estimation 34

7.2 Kaplan-Meier Estimator 36
7.3 Censored MLE Estimator 37

8 A learning model - formulation and simulation **38**
8.1 Simulation results 39

9 Conclusion **67**

10 Appendix - Simulation code **69**
10.1 LRNSYS.M 69
10.2 CLN_QUE.M 72
10.3 CLIENT_Z.M 74
10.4 UPDT_CLI.M 76
10.5 Tz_MLE.M 77
10.6 Tz_APRME.M 78
10.7 UPDT_PT.M 79

List of Figures

1 *estimation of the waiting time distribution customers 1-4* 40
2 *estimation of the waiting time distribution customers 5-8* 41
3 *estimation of the waiting time customers 1-4* 42
4 *estimation of the waiting time customers 5-8* 43
5 *estimation of the waiting time 1-4* 46
6 *estimation of the waiting time 5-8* 47
7 *estimation of the waiting time 9-11* 48

8 *estimation of the waiting time distribution* 49

9 *Kaplan-Meier estimation of the waiting time distribution, customers 1-4* . . 52

10 *Kaplan-Meier estimation of the waiting time distribution, customers 5-8* . . 53

11 *Kaplan-Meier estimation of the waiting time distribution, customers 9-12* . 54

12 *Kaplan-Meier estimation of the waiting time distribution, customers 13-16* 55

13 *Kaplan-Meier estimation of the waiting time distribution, customers 17-18* 56

14 *Kaplan-Meier estimation of the waiting time, customers 1-4* 57

15 *Kaplan-Meier estimation of the waiting time, customers 5-8* 58

16 *Kaplan-Meier estimation of the waiting time, customers 9-12* 59

17 *Kaplan-Meier estimation of the waiting time, customers 13-16* 60

18 *Kaplan-Meier estimation of the waiting time, customers 17-18* 61

19 *Censored MLE estimation of the waiting time 1-4* 62

20 *Censored MLE estimation of the waiting time 5-8* 63

21 *Censored MLE estimation of the waiting time 9-12* 64

22 *Censored MLE estimation of the waiting time 13-14* 65

23 *The theoretic and real patience function* 65

1 Abstract

We consider the behavior of a queueing system, as an outcome of its customers' patience. The patience function (the probability for a randomly selected customer to stay in the queue t seconds after arrival) is assumed to be related to the system's performance, as perceived by its customers, thus implying a mutual dependence between the two. This setup is a game (in the game theoretic sense), whose equilibrium (existence and uniqueness) is explored. The system is modeled as an M/M/m queue with abandonment, where the queue is invisible (i.e. the queue length is not revealed to the waiting customers). The only evidence available to a customer in such a queue is whether or not he was served, and how long he has waited. Since it is very natural for people to generalize and then to deduce, it is reasonable to believe that customers use their collected evidence to estimate the mean waiting time (generalization) and then, this parameter is used for deduction (e.g. decision of when to hang up). Consequently, we have selected models, which consider the estimated mean waiting time, x, as the main factor influencing customers' expectations and behavior. We study two classes of functions, which describe the customers patience. The first is the class of all patience functions decreasing in x, and the second consists of all patience functions increasing in x, in a linearly bounded way. Additionally, we suggest a rational decision model, used to describe the decision process performed by each waiting customer, as an individual. The customers are assumed to believe in an exponentially distributed waiting time, and also to be affected by a convex increasing waiting cost function. The aggregate behavior of the rational customers is then related to the above classes of customers patience functions. Both the rational model as well as the explored patience functions sets assume minimal knowledge of the system performance and state, relegated through a single parameter - the estimated mean waiting time x. We show for the class of functions decreasing in x that there exists an equilibrium to the system and that this equilibrium is unique. This result is then extended to the case of patience functions which are increasing in x, in a linearly bounded way. In particular, we obtain existence and uniqueness of

equilibrium for the rational decision model by showing that the aggregate behavior of customers according to this decision rule yields patience functions within the above classes. We finally consider in simulation a dynamic model of learning for the system, where the customers use their past experience to estimate the mean waiting time in the system using a censored sampling estimator (both the Censored MLE and the Kaplan-Meier estimator are explored). The estimated mean waiting time is exploited by the rational decision model, to determine whether or not to leave the queue at each given time. Convergence of the system to the theoretically anticipated equilibrium is demonstrated for the Kaplan-Meier estimator, and the biassed result recevied upon using the Censored MLE estimator is analytically justified. This work was highly motivated by the need to provide means with which administrative decisions can be performed. The models developed herein can be used for the prediction of customers behavior in invisible queues, and in particular their adaptation to changes in system performance. Thus, providing means for the design of more cost-effective systems, by allowing for educated decisions of whether to add, improve or eliminate servers.

2 Notations list

v - offered waiting time.

x - estimated mean offered waiting time.

$\bar{G}(x,t)$ - patience function, the probability for a randomly selected customer to stay in the queue until time t, given that the estimated waiting time is x.

λ - arrival rate.

$\frac{1}{\mu}$ - mean service time.

p_j - stationary probability for j occupied servers.

$F'(x,t)$ - offered waiting time density function.

$H_z(x,t)$ - offered waiting time hazard rate function.

$E_x[v]$ - mean offered waiting time, calculated according to the density function $F'(x,t)$.

$T_z(x)$ - abandonment time of a z type customer with estimated waiting time x.

P_z - customers types distribution.

w_z - relative weight of a z type customer in the group of all customers types.

3 Introduction

3.1 Background and Motivation

This research deals with the analysis of customers' abandonment from invisible queues. It is highly motivated by the real need of understanding and addressing customers' expectations from services with which invisible queues are inherently embodied. The most evident example for such services is telephone call-centers, which have become profit centers for many businesses. In particular telephone call-centers, which do not reveal the place of the waiting customer in the queue, hence invisible queues. According to the Direct Marketing Association [5] direct sales and marketing via call centers accounted for $308 billion or 44.6 percent of total business-to-business sales in 1999 (overall sales via call centers: $538.3 billion). In the last years, this market has proliferated even further mainly due to the Internet. The web clearly provides a new frontier for call centers, which have the potential of humanizing the web-experience, and supporting on-line consumers with the comfort and security they need, rendering "call centers" into "virtual customer care centers". By using call centers to address and enhance customers' satisfaction and security, companies manage to retain their customer-base, keeping in mind that a customer satisfied, in many cases, equates to future dollars earned.

It comes down to serving user's needs, and in today's world: "time means money", has been automatically translated to "reducing the time customers spend in queues". But our goal is to serve "customers' expectations", and since customers are adapting to the current quality of service, and rate their satisfaction according to these expectations, the question raised is: What are the basic elements, which build those expectations? In the following we address these issues, in order to model customers behavior in invisible queues, and especially their adaptation to changes in system performance, hence providing means for the design of more cost-effective systems.

3.2 Assumptions and Results

The only evidence observable by a person in an invisible queue (e.g. telephone) is whether or not he was served, and how long he has waited. We strongly believe that it is within human nature to base one's expectations upon perception of prior experience. Therefore, we persist that this is the case in the invisible queue case, and that the limited evidence available is used to build one's expectations of the service. Furthermore, since it is very natural for humans to generalize and then to deduce, it is reasonable to believe that customers use their collected evidence to estimate their mean waiting time (generalization) and then, this parameter is used for deduction (e.g. decision of when to hang up). Consequently, we've selected, in the following, a model, which considers the estimated mean waiting time as the main factor influencing customers' expectations and behavior.

We model the behavior of customers, using two complementary approaches. Both approaches summarize a customer's expectations through a single parameter, the average waiting time for service. The first approach, a more general one, refers to the abandonment behavior of all customers, expressed by a global patience function with general characteristics. This gives us tools to analyze empirical data, or even specific decision rules. The second model applies an individual "rational decision rule": Each and every waiting customer weights the possible benefit from getting the service against the cost of waiting, and the possibility of not getting a service at all. The optimal abandonment time is then drawn out of the individual knowledge about the queue.

Based on the general attributes of the patience function, we show that a global patience function, which decreases with x, has a unique equilibrium. We also refer to increasing patience function, in particular to a set of increasing patience functions with a bounded increase rate. We show that if the maximum rate of the increase rate of the patience function is less than the increase rate of x, then there exists a unique equilibrium.

In the rational abandonment decision model, an individual decision is derived from a comparison of the waiting time hazard rate function with a waiting cost function. We

suggest two types of cost functions, one which does not depend on x at all, and the other with an additive dependence. By assuming that the customers believe in an exponential distributed waiting time, the hazard rate may be easily derived. We show that the first type of cost function (with dependence on t alone) yields a decreasing (in x) global patience function, which guarantees te existence of a unique equilibrium. The second cost function corresponds to the bounded increasing patience, therefore, if the equilibrium exists - it is unique.

3.3 A learning model - formulation and simulation

We suggest a simple learning model based on the following behavioral assumption - the estimated mean waiting time is the main factor influencing customers' patience. According to the model customers learn about the queue through their experience of entering the service or abandoning the queue. Every customer decides whether or when to abandon the queue after estimating the expected waiting time by using his previous experience. The abandonment experiences provide only the lower bound on the waiting time till service, therefore a censored estimator is required. Two estimators are considered: Kaplan-Meier estimator, and censored MLE. Although the first estimator is a more accurate one, (since we do not assume any underlying distribution) it is not very likely that customers will use such a complicated estimator. The censored MLE estimator, on the other hand, is a more intuitive estimator, but it assumes an exponential distribution for the waiting time.

The performance of the system with the sggested learning model is examined via simulation, and its convergence to the anticipated equilibrium is demonstrated. The simulation is of an M/M/m + G queue. Customers enter the queue and wait until one of two events takes place: (1) they are admitted into service (2) they lose their patience and decide to abandon. When the customers use the Kaplan-Meier estimator the estimated mean waiting time after convergence is very close to the theoretic equilibrium result. The use of the Censored MLE yields a biased equilibrium, which is analytically justified.

Another interesting result is the convergence of the mean values in a system with two types of customers, who are using the censored MLE estimator. The distinction between the groups is via the group's patience. Customers from one group are willing to wait a longer period of time than the customers in the second group, given the same expected waiting time. The two groups are clearly distinct in their final estimated values, where customers with less patience reach higher values of estimated waiting time. This is a result of the incorrect assumption of the censored MLE - the exponential distribution of the waiting time. We show analytically how under the false assumption we get similiar results to the simulation results. Simulations of two customers types using the Kaplan-Meier estimator yielde similiar results for both types, which coincide with the theoretic value of the equilibrium mean waiting time.

3.4 Contents

The following section presents an overview of related research. In section 5 we describe the model suggested for customers' behavior. The section begins with modeling the system and analyzing it using the results of Baccelli and Hebuterne [1], with a general patience function $\bar{G}(t)$. The section further details the rational decision model, by derivation of the optimal abandonment rule, assuming a constant hazard rate function (exponentially distributed waiting time belief). Additionaly, an overview of psychological references is provided, to explain the characterization of the cost function to be used. In section 6 we present conditions for uniqueness and existence of the partially consistent equilibrium, based on the assumptions from section 5. Having established the characterization of the equilibrium, we demonstrate, in section 8, our results through simulation. Finally, a summary and concluding remarks are found in section 9.

4 Related research

Facing a problem, which involves formulation of customers' behavior, one must remember the limitations of the analytic and computation tools available. In other words, taking into consideration each and every decision of all customers at all times, is not feasible. The following articles address the probelm of modeling behavior of customers, who are waiting on the telephone line, both from the psychological and the mathematical points of view.

4.1 Related work on queueing systems

A model which describes the individual rational decision rule of waiting customers is proposed by Mandelbaum and Shimkin in [19]. In their article (which is highly related to this work) they assumed an invisible queue, but the consistency assumption imposed required that customers have complete knowledge of the offered waiting time distribution function, rather than a partial knowledge like the estimated mean waiting time. The decision rule compares the hazard rate function and a constant marginal cost-benefit ratio (or linear cost-benefit ratio). The simple M/M/m + G model results in zero or infinite patience, which does not seem to fit the intuitive idea of customer behavior. An M/M/m(q) + G model adds the probability of (1-q) for a fault state in the system, in which the customer does not obtain service no matter how long he waits. This model results in a non-trivial patience profile, with a unique equilibrium.

Another paper related to the rational decision to abandon an invisible M/M/1 queue is the work of Hassin and Haviv [10], where homogeneous customers (identical cost function and value of service), perform individual decisions, whether to join the queue and when to abandon. A customer is assumed to join the queue with probability p, and with a deadline - T for reneging (since he does not get any new information while waiting). The decision of what deadline to choose is based on a linear cost function and a fixed value of service. A unique Nash equilibrium is shown to exist for the pair (p, T). The work of Haviv and Ritov [11] establishes the conditions for a unique Nash equilibrium in an M/M/1+G queueing

system. The customers are homogeneous and with a convex cost function.

Osuna [20] provided a model for explaining the cost of waiting. The main contribution of this paper is the consideration of the psychological cost, along with the economic cost. During the wait, a process of building up stress takes place. The higher the uncertainty the higher the stress. The stress accumulates to be the psychological cost of the waiting. Osuna also refered to the expectations issue by showing the consequences of providing the customer with information regarding the waiting period. [24] proposes a generalization to Osuna's work by eliminating the the need for the constraints of a bounded $G(t)$, which does not have common discontinuities with $F(t)$, in the monotonicity theorem.

Palm [21] suggested a parametric model to characterize the inconvenience caused to customers waiting in a congested system. The paper begins by describing the inconvenience function as an integral over the irritaion of the waiting customer, which is intuitively assumed to be an increasing function in t. Then, Palm assumes that the probability for a customer to abandon during the time interval $[t, t + dt]$, given he has been waiting till t, equals the corresponding irritation in the interval. Taking advantage of some unconnected exchanges in the Stockholm area allowed to measure when customers hang up without getting the service. Thus supplying Palm a way to estimate the values of theS model's parameters.

4.2 Psychology related articles

The following papers are related to the psychological aspects of waiting. We will refer to these papers in the section 5.3.2, where we explain the characteristics of the waiting cost function (non-linear increasing).

Thierry in [27] raises two central issues: (1) Approaching waiting lines as goal-oriented settings. (2) The individual experience at all time in regarding to the goal. The aim of the study was to assess time-integration processes as a function of the subjective importance of the goal. The connection between the subjective importance of the goal and the way

passing of time is experienced by the customer was tested in a field observation. The observed queue was a queue to an exhibition, the theme of which was the work of Gauguin. Participants were categorized as highly or lowly goal oriented according to their preferences of the painters from a list of 20 painters of Gauguin's period (highly goal oriented - chose Gauguin as one of their three preferred painters). All the participants answered three questions: (1) what is their current mood, (2) what is their estimation for the duration of waiting they still have ahead, (3) what is their evaluation of the number of people ahead of them. The results showed that there was no significant difference between the two categories in mood, but participants with high goal orientation estimated their distance from the goal to be shorter then the distance estimated by the low orientation participants. Testing the data within each category revealed a strong connection in time spent in the queue and the current mood (the longer they waited or the greater the distance, the less pleasant was the mood), only for the low oriented category. The low oriented group also tended to underestimate the remaining waiting time when being far away from the goal and overestimate it when being closer to the goal. The high oriented group relied heavily on the distance from the goal (and not on the time they already spent in the queue) in order to estimate the remaining waiting period, whereas low oriented relied also on the time spent waiting.

In [7] Friedman and Friedman state that the cost of waiting is different for different people, and with this idea they propose the waiting line segmentation. They begin by segmenting the customers into two groups: willing to pay in order to shorten the waiting time, and willing to wait. Servers are assigned to each group in a way that paying customers get a higher rate of service. We can see a possible game arising between the size of the payment and the difference in waiting time (low payment leads to many paying customers and so to longer waiting time). The steady-state results proved that with this method we can increase firm's revenue, and improve customers satisfaction. (The article includes graphs and data of customers' choices due to given costs of service).

Larson [17] reviews several different factors, which play a major part in influencing peo-

ple's experiences while waiting in queues. Among these factors one finds social injustice (breaking the unwritten law of FIFO), queueing environment (distracting the customers attention from the waiting itself), delay feedback (information about the delay estimation). The author concludes with two important points. One point is supported by a research, which was conducted in IBM. This research shows how the productivity of a time-shared computer users varies nonlinearly with the delay, characterized by an elbow in the productivity vs. computer response time graph. The other point is the nonlinearity of the disutility function with the queueing delay, caused by the annoyance that waiting customers feel. A short report regarding the latter issue is added to the article. This report states that almost all disutility functions where found to be nonlinear with queueing delays.

Dube, Schmitt, and Leclerc [6] describe a field experiment, investigating mood changes as a reaction to delays of students waiting for their TA to return to class. Three types of delays were compared: preprocess, postprocess, and in-process. The results showed that subjects, who experienced pre and postprocess delays, expressed higher degrees of negative feelings than the in-process delay and the control groups (e.g. no delay).

Taylor has two relevant papers. In [25] she explores customer reaction to a service delay by assessing the relationship between the delays and evaluations of service, and integrating those relations into a conceptual model. Waiting times are categorized as pre, in, and post-process. Pre-process is also divided to pre-schedule (e.g. arriving before the scheduled event, therefore waiting the pre-schedule period along with the delay), delays (post-schedule), and queue waits (no scheduled event but FIFO). This article focuses on examining the degrees of anger and uncertainty (elements of the evaluation) caused by delays. The results showed a strong relation between the customers anger and uncertainty and lower evaluations of service. The waiting time estimation of the customers was also influenced by anger and uncertainty, the stronger those feelings were the longer the time was perceived.

The other article written together with Claxton is [26]. Although occupied with similar questions, it concentrates on alterations of stable service: What will happen after chang-

ing one service attribute to the other attributes? Will service evaluations and importance weightings change? (The questions' motivation is checking the stability of a linear compensatory model). Delay was found as affecting respondents' moods (significant difference between delayed and non-delayed). It was also found that the importance of punctuality, on the overall evaluation of service, increased due to a negative experience.

Diekmann, Jungbauer-Gans, Krassnig and Lorenz [4] describe another field experiment, which examined the relation between drivers' characteristics and the type of the cars they have. The experiment involved a blocking car and an observation of the blocked driver reaction latency. There was also a use of Kaplan-Meier technique in the cases where the drivers' only reaction was leaving the blocked lane (no honking or beaming). The results showed that drivers of higher social status cars tend to be more aggressive than those of lower status cars. The estimated survival functions of response time consisted of an "elbow" after the first few seconds, which means high rate of response till the "elbow" time and low rate of response later on.

Gail and Scott present in [8] a field observation, which took place in two separate supermarkets. The observation purpose was to determine the effects of objective waiting time, perceived waiting time, and serve time on customer satisfaction with the server and the store. The results confirmed previous findings that shorter perceived time leads to more satisfaction of the customer with the server (the satisfaction with the store remains uninfluenced). The perceived waiting time was greater than or comparable to objective waiting time. Another important conclusion claims that there are situations when customers will be more satisfied with longer period of waiting. These situations are, for example, the cases when checkers spend time socializing with the customers.

Maister introduces 2 propositions about service encounters [18]: 1. Satisfaction equals perception minus expectation. 2. It's hard to play catch-up ball. The early stages of the service are the important ones, so we must concentrate in improving this period. The author also introduces his 8 propositions about the psychology of queues:

1. Unoccupied time feels longer than occupied time.

2. Pre-process waits feel longer than in-process waits.
3. Anxiety makes waits seem longer.
4. Uncertain waits are longer than known, finite waits.
5. Unexplained waits are longer than explained waits.
6. Unfair waits are longer than equitable waits.
7. The more valuable the service, the longer I will wait.
8. Solo waits feels longer than group waiting.

Hueter and Swart describe in [13] the work that was done for the Taco Bell Corporation in order to improve its efficiency and cost-effectiveness. The writers investigated all the aspects of the labor-management system, but a specific point of their work is of great importance for modeling customers' patience in queues. They performed a study, which assessed when customers would be likely to leave the queue (due to perceiving the wait as excessive). According to the results, customers perceived the first 5 minutes of the wait as approximately 2 minutes. At about 5 minutes the graph is characterized by an "elbow", which translates the loss of patience experienced by customers.

A somewhat incomplete work on the shopping behavior is presented by Hornik in [12]. The author suggests 4 hypotheses, but the study doesn't refer to the last and very interesting one ("The frequency of shopping, in addition to personal characteristics, might influence individual perceptions of waiting time"). The results of this paper showed that customers tend to overestimate waiting time during the waiting period.

5 Model Formulation

Our model combines two points of view, that of the analyst and that of the common customer. The analyst tries to understand waiting or abandonment decisions through the knowledge of the actual waiting time distribution. This case was studied in [1]. On the other hand, the customer has only partial observations of the scenario: the history of his previous trials using the service, and maybe some information from other customers. In the following we consider two separate, though related, approaches for the characterization of the customers' behavior. One approach perceives the decision process as an outcome of an optimization, combining prior knowledge with cost-benefit ratio functions. The other uses an aggregate patience function to describe the customers' decisions, representing the probability that a randomly chosen customer stays in the queue at time t. Within both frameworks, the entire knowledge, available to the customers about the system, is summarised by a single parameter: the estimated mean offered waiting time - x (aka. the estimated waiting time). In general, x may be different from the calculated mean waiting time, since it reflects only the knowledge available to individuals rather than complete understanding of the system. Hence, we introduce an assumption for this partial knowledge about the system:

Definition 1 *The partial consistency assumption is defined as the case where the estimated and calculated waiting time are* equal

5.1 System model

We will assume throughout this paper that our system is an M/M/m queue with Poisson arrival rate - λ, an exponential service time with mean value of $1/\mu$, and m servers. Let $G(x,t)$ denote the probability for a randomly selected customer to abandon the queue until time t, provided that her estimated waiting time is x. Then let $\bar{G}(x,t) = 1 - G(x,t)$ be the *patience function.* The customers in the system are allowed to abandon the queue at

any time (prior to receiving service), and the global abandonment probability is expressed through the function $\bar{G}(x,t)$. We assume that $m\mu > \lambda\bar{G}(x,\infty)$ for every x. Consequently, as in [1], the density function of the offered waiting time is:

$$F'(x,t) = \lambda p_{m-1} \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds) \ , \ t \geq 0 \ ,$$

Where p_{m-1} is the stationary probability for having exactly $m-1$ occupied servers.
The normalization condition is:

$$\sum_{j=0}^{m-1} p_j + \int_0^{\infty} F'(x,t)dt = 1 \ , \quad p_j = \left(\frac{\lambda}{\mu}\right)^j \frac{1}{j!} p_0 \ .$$

Here p_j is the stationary probability for j occupied servers, $j = 0, 1..., m$.
Therefore, the density function for an M/M/m+G may be written as:

$$F'(x,t) = \frac{\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)}{\frac{K_m}{\lambda} + \int_0^{\infty} \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt} \tag{5.1}$$

$$K_m = \sum_{j=0}^{m-1} \frac{(m-1)!}{j!} \left(\frac{\lambda}{\mu}\right)^{j-m+1} \tag{5.2}$$

In the following sections we suggest two approaches for characterizing customers abandonment behavior. The first approach refers to the global patience function, and draws some guidelines to the attributes of the calculated mean waiting time in the system. The other, optimizes individual utility function, as in [19], but under less demanding assumptions. Observe that a customer waiting in a queue has expectations from the serving system regarding the time she is about to waste during the wait period. In particular, our main assumption, with respect to these expectations, is that (in the case of invisible queue) they are highly dependent on the customer's estimated mean waiting time. This estimated mean

is formed mainly by the learning experience of the customer, during her past experiences in the system, but may also be influenced by hearsay from other customers.

5.2 The patience function

Let $\bar{G}(x,t)$ be a general patience function. (Recall that this function represents the probability for a randomly chosen customer to keep on waiting at time t, given her estimated waiting time x). If $\bar{G}(x,t)$ is decreasing in x, it means that the probability to find a randomly selected customer in the queue at time t is smaller for larger x. Equivalently, if the customers estimated the waiting time as longer, they would be willing to wait less, and thus, lead to a shorter waiting time. In the next proposition we formalize this statement and prove it using stochastic order techniques.

Proposition 5.1 *Let $\bar{G}(x,t)$ be the patience function, and $F(x,t)$ the corresponding offered waiting time distribution function, defined through 5.1. Assume that $\bar{G}(x,t)$ is a decreasing function in x for every t. If W_1, W_2 are random variables with the distributions $F(x_1,t), F(x_2,t)$ respectively and $x_1 > x_2$, then $W_1 \leq_{st} W_2$, in particular $E[W_1] \leq E[W_2]$.*

Proof: The proposition states, in other words, that the calculated waiting time is a stochastically decreasing function of x. Let us define the following function, in order to simplify the expressions we are about to use:

$$J(x,t) := \int_0^t (m\mu - \lambda \bar{G}(x,s))ds \tag{5.3}$$

$$\triangle \bar{G}_\delta(x,s) := \bar{G}(x,s) - \bar{G}(x+\delta,s) \tag{5.4}$$

Since $\bar{G}(x,t)$ is a decreasing function then for $x_1 > x_2$ there is a function $\triangle \bar{G}_\delta(x,s) \geq 0$, which satisfies the following equation:

$$J(x_2,t) = J(x_1,t) - \lambda \int_0^t \triangle \bar{G}_{x_1-x_2}(x_2,s)ds \tag{5.5}$$

The hazard rate functions of the two random variables are:

$$H_i(t) = \frac{F_i'(t)}{\bar{F}_i(t)} = \frac{\lambda p_{m-1} exp(-J(x_i,t))}{\lambda p_{m-1} \int_t^\infty exp(-J(x_i,v))dv} = \frac{exp(-J(x_i,t))}{\int_t^\infty exp(-J(x_i,v))dv} \tag{5.6}$$

where i is the index of the distribution. By substituting the function with the expression from (5.5), we get:

$$\begin{aligned} H_2(t) &= \frac{exp(-J(x_1,t))exp(\lambda \int_0^t \triangle \bar{G}_\delta(x_2,s)ds)}{\int_t^\infty \left[exp(-J(x_1,v))exp(\lambda \int_0^v \triangle \bar{G}_\delta(x_2,s)ds)\right] dv} = \\ &= \frac{exp(-J(x_1,t))}{\int_t^\infty \left[exp(-J(x_1,v))exp(\lambda \int_0^v \triangle \bar{G}_\delta(x_2,s)ds)exp(-\lambda \int_0^t \triangle \bar{G}_\delta(x_2,s)ds)\right] dv} \end{aligned}$$

Since,

$$exp(\lambda \int_0^v \triangle \bar{G}(s)ds)exp(-\lambda \int_0^t \triangle \bar{G}(s)ds) \geq 1$$

We have:

$$H_2(t) \leq H_1(t)$$

The inequality $H_2(t) \leq H_1(t)$ means that W_1 is smaller than W_2 in the hazard rate order, or $W_1 \leq_{hr} W_2$. According to theorem 1.B.1 from [23] this implies that $W_1 \leq_{st} W_2$. It is easily seen that the last stochastic relation leads to $E[W_1] \leq E[W_2]$.

□

Similarly to proposition 5.1, the following states that If $\bar{G}(x,t)$ is increasing in x, then the probability to find a randomly selected customer in the queue at time t is bigger for larger x.

Proposition 5.2 *Let $\bar{G}(x,t)$ be the patience function, and $F(x,t)$ be the corresponding offered waiting time distribution function, defined through 5.1. Assume that $\bar{G}(x,t)$ is a increasing function in x for every t. If W_1, W_2 are random variables with the distributions $F(x_1,t), F(x_2,t)$ respectively and $x_1 > x_2$, then $W_1 \geq_{st} W_2$, in particular $E[W_1] \geq E[W_2]$.*

Proof: The proof is similar to that of proposition 5.1, with the corresponding relations.

□

We now continue to explore a specific group of patience functions. We will calculate

the *objective* mean waiting time in the system, using the density function from (5.1) for the M/M/m+G queue. Denote $E_x[v]$ - the calculated (*objective*) mean waiting time, and $E_x[v|v > 0]$ - the calculated waiting time given that there exists a waiting period. In proposition 5.3 below we show that $E_x[v|v > 0]$ is increasing with x with a less than 1 slope, for the simple case of a linear shift of $\bar{G}(x,t)$. This result is extended for the general case of a shift - $shft(x)$, which satisfy $0 \leq \frac{\partial}{\partial x} shft(x) \leq 1$, in proposition 5.4. Let us begin by explaining the motivation for working with this group of functions. Consider, for simplicity, a finite number of customers of types z, in which the abandonment time function is:

$$T_z(x) = x + \beta(z), \tag{5.7}$$

where z is the customers type, distributed according to P_Z, over N different types. Assuming that all customers stabilized on the same estimated mean waiting time, then the patience distribution $\bar{G}_z(x,t)$ for one type of customers would be the following step function:

$$\bar{G}_z(x,t) = \begin{cases} 1 \text{ for } t \leq x + \beta(z) \\ 0 \text{ for } t \geq x + \beta(z) \end{cases} \tag{5.8}$$

Denote w_z the relative weight of customer type z, according to the distribution P_z. Then,

$$\sum_{z=0}^{N} w_z = 1$$

and the patience function of all customers is given by:

$$\bar{G}(x,t) = \sum_{z=0}^{N} w_z \bar{G}_z(x,t). \tag{5.9}$$

We see that an increase in x results in a shift to the right of the function $\bar{G}(x,t)$.

$$\bar{G}(x,t) = \bar{G}(t-x) \tag{5.10}$$

The same shift would appear in the continuous case:

$$P(T_z(x) > t) = P(x + \beta(z) > t) = P(\beta(z) > t - x) = \bar{G}(t - x)$$

for some distribution function G. (In fact, G is the distribution function of the random variable $\beta(z)$, where z is distributd according to P_Z.)

The calculated (*objective*) mean waiting time according to the density function from (5.1) is:

$$E_x[v] = \frac{\int_0^\infty t \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt}{\frac{K_m}{\lambda} + \int_0^\infty \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt} \tag{5.11}$$

The last expression allows for the possibility of being admitted into service immediately, or zero waiting time. We believe that it is more natural to separate the two events. If a customer does not need to wait for service, there is no reason to address an hypothetical patience function. However, if a waiting period is required, then the customer's patience is examined. Therefore, we will refer to $E_x[v|v > 0]$, as the main objective element for customers to evaluate their expectations from a waiting period. Furthermore, we will assume later that customers maintain an *exponential distribution of waiting time* belief. This belief does not assign a positive probability to the event $v = 0$.

The conditional calculated waiting time is:

$$E_x[v|v > 0] = \frac{\int_0^\infty t \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt}{\int_0^\infty \exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt} \tag{5.12}$$

Proposition 5.3 *Assume that $\bar{G}(x,t)$ - the patience function, is linearly shifted by x, as in (5.10). Also assume the partial consistency assumption (definition 1). Then the coditional calculated waiting time (5.12) is an increasing function in x, which satisfies:*
$\frac{\partial}{\partial x}E_x[v|v > 0] \leq 1$

Proof: We have already proven that if $\bar{G}(x,t)$ increases with x, then the calculated mean also increases with x (see (5.2)). We now turn to further analyze $\frac{\partial}{\partial x}E_x[v]$. The relation between the derivatives according to x and t, using the characteristic from (5.10) is:

$$\frac{\partial \bar{G}(t-x)}{\partial x} = -\frac{\partial \bar{G}(t-x)}{\partial t} \tag{5.13}$$

and therefore:

$$\int_0^t \frac{\partial \bar{G}(s-x)}{\partial x} ds = \bar{G}(x,0) - \bar{G}(x,t) = 1 - \bar{G}(x,t)$$

Using the last relation, the derivative of the exponent in the expression of $E_x[v]$ according to x is:

$$\frac{\partial}{\partial x}\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds) = \left(\lambda - m\mu + m\mu - \lambda\bar{G}(x,t)\right)\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)$$

and according to t:

$$\frac{\partial}{\partial t}\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds) = \left(\lambda\bar{G}(x,t) - m\mu\right)\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)$$

Combining the above:

$$\frac{\partial}{\partial x}\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds) = (\lambda - m\mu)\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds) -$$
$$-\frac{\partial}{\partial t}\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)$$

Denote A and B as follows:

$$\text{numerator}\,(E_x[v]) \equiv \text{A} = \int_0^\infty t\exp(-\int_0^t \left(m\mu - \lambda\bar{G}(x,s)\right) ds)dt$$

$$\text{denominator } (E_x[v]) \equiv B = \frac{K_m}{\lambda} + \int_0^\infty \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt$$

$$\frac{\partial A}{\partial x} = (\lambda - m\mu) \int_0^\infty t \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt - \int_0^\infty t \frac{\partial}{\partial t} \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt =$$
$$= (\lambda - m\mu) A - t \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) \Big|_0^\infty + \int_0^\infty \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt =$$
$$= (\lambda - m\mu) A + B - \frac{K_m}{\lambda}$$

$$\frac{\partial B}{\partial x} = (\lambda - m\mu) \int_0^\infty \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt - \int_0^\infty \frac{\partial}{\partial t} \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) dt =$$
$$= (\lambda - m\mu) \left(B - \frac{1}{\lambda}\right) - \exp(-\int_0^t \left(m\mu - \lambda \bar{G}(x,s)\right) ds) \Big|_0^\infty =$$
$$= (\lambda - m\mu) \left(B - \frac{K_m}{\lambda}\right) + 1$$

$$\frac{\partial}{\partial x} E_x[v] = \frac{B \frac{\partial}{\partial x} A - A \frac{\partial}{\partial x} B}{B^2} = 1 - \frac{\frac{K_m}{\lambda} B + A - (\lambda - m\mu) \frac{K_m}{\lambda} A}{B^2}$$

We know that for the conditional calculated waiting time $K_m = 0$:

$$\frac{\partial}{\partial x} E_x[v | v > 0] = \frac{B \frac{\partial}{\partial x} A - A \frac{\partial}{\partial x} B}{B^2} = 1 - \frac{A}{B^2} \leq 1$$

□

Remark: It is easy to see the the last proposition holds for the calculated waiting time $E_x[v]$ of an M/M/1+G, where $K_m = 1$, and for every K_m if $\lambda \leq m\mu$.

Proposition 5.4 *Define $shft(x)$ as the shift function of $\bar{G}(x,t)$, and suppose that $shft(x)$ is non-decreasing with $\frac{d(shft(x))}{dx} < 1$. Then under the partial consistency assumption, the conditional calculated waiting time is increasing in x, and satisfies $\frac{\partial}{\partial x} E_x[v | v > 0] < 1$*

Proof: The general shift is now:

$$\bar{G}(x,t) = \bar{G}(t - shft(x))$$

Expression 5.13 takes the more general form:

$$\frac{\partial \bar{G}(t - shft(x))}{\partial shft(x)} = -\frac{\partial \bar{G}(t-x)}{\partial t}$$

or:

$$\frac{\partial \bar{G}(t - shft(x))}{\partial shft(x)} \frac{d\left(shft(x)\right)}{dx} = -\frac{\partial \bar{G}(t-x)}{\partial t} \frac{d\left(shft(x)\right)}{dx}$$

The derivatives of A and B are now:

$$\frac{\partial A}{\partial x} = \left[(\lambda - m\mu)\, A + B\right] \frac{d\left(shft(x)\right)}{dx}$$

$$\frac{\partial B}{\partial x} = \left[(\lambda - \mu)\, B + 1\right] \frac{d\left(shft(x)\right)}{dx}$$

Hence, the derivative $\frac{\partial}{\partial x} E_x[v]$ is:

$$\frac{\partial}{\partial x} E_x[v] = \frac{B\frac{\partial}{\partial x}A - A\frac{\partial}{\partial x}B}{B^2} = \left[1 - \frac{A}{B^2}\right] \frac{d\left(shft(x)\right)}{dx}$$

$$0 < \frac{\partial}{\partial x} E_x[v|v > 0] \leq 1$$

□

5.3 Rational abandonment

5.3.1 The utility function

The rational abandonment decision is based on the individual utility function of the customers. We distinguish between different types of customers according to their decision model parameters. Let $z \in Z$ denote the type, with Z the set of possible types. A customer of type z will be characterized by the following elements:

(*i*) r_z, the service utility, assumed to be positive.

(*ii*) $c_z(t)$, the marginal cost function of waiting, assumed positive and increasing, as explained later on.

(*iii*) P_Z, a probability distribution over the set of customer types Z. The type z of each customer is randomly chosen according to P_Z, independently across customers.

Along with the type parameters, all customers are assumed to believe in an exponentially distributed waiting time. The parameter of the distribution - *estimated mean waiting time*, can be different for each customer without relation to type of customer.
Define the *cost-benefit ratio*function: $\gamma_z(t) := c_z(t)/r_z$. We will see that this function dictates the abandonment time of the customers in each type group, given the same estimated mean waiting time. Denote the cost function:

$$C_z(t) = \int_0^t c_z(s)ds.$$

The utility function, at any possible abandonment time $T \geq 0$, accounts for the probability to be admitted into service, and compares the service utility against the total waiting cost, as follows:

$$\begin{aligned} U_z(T) &= E_z(r_z 1\{T \geq V\} - C_z(\min\{V,T\})) \\ &= \int_{0-}^{T} [r_z - C_z(t)]dF_z(v) - C_z(T)\bar{F}_z(T)\,, \end{aligned} \tag{5.14}$$

where E_z is the mean according to the subjective exponential probability, and V is the offered waiting time. The optimal abandonment time is the one that maximizes the utility function. Differentiating the utility function (5.14) with respect to $T > 0$ gives

$$\begin{aligned} U_z'(T) &= [r_z - C_z(T)]F_z'(T) - c_z(T)\bar{F}_z(T) + C_z(T)F_z'(T) \\ &= r_z F_z'(T) - c_z(T)\bar{F}_z(T) \end{aligned} \tag{5.15}$$

Since $r_z > 0$ by assumption, when $\bar{F}_z(T) > 0$ this may be written in the following way:

$$\begin{aligned} U_z'(T) &= r_z\bar{F}_z(T)[F_z'(T)/\bar{F}_z(T) - \gamma_z(T)] \\ &= r_z\bar{F}_z(T)[H_z(T) - \gamma_z(T)] \end{aligned} \tag{5.16}$$

where H_z is the hazard rate function associated with the offered waiting time distribution, namely

$$H_z(t) := F_z'(t)/\bar{F}_z(t)\,, \quad t > 0.$$

$U_z'(T) = 0$, can now be simply stated as

$$H_z(T) = \gamma_z(T)\,. \tag{5.17}$$

Recall the assumption regarding customers' belief from the beginning of this subsection. We characterized the customers with an exponential distributed waiting time belief. The hazard rate function of an exponential distribution is:

$$H_z = \frac{F'}{\bar{F}} = \frac{\mu_z exp(-\mu_z t)}{exp(-\mu_z t)} = \mu_z \tag{5.18}$$

H_z is constant in time, and equals the inverse of the mean waiting time. Hence, (5.17) reduces to:

$$\gamma_z(T) = \mu_z.$$

5.3.2 The cost function

The rational decision to abandon, as established in the last subsection, is based on an optimization process which is affected by the customers' limited knowledge about the system (gathered through the waiting experience) as well as their waiting cost function. The purpose of the waiting cost function in the process, as we define it, is to combine all the elements, which influence a customer to abandon a queue (or not to enter at all). How does this waiting cost function look like? In order to answer this question let us elaborate on the psychological related research presented in section 4.

The most related assessment of customers feelings while waiting in a queue appears in the work of Carmon and Kahneman [3], who tested the momentary affect on waiting customers. They showed clear results of dependence between the advances in an invisible queue and customers' feelings. After every advance in the queue the affect meter jumped up (positive feelings), and then declined slowly till the next advance. We believe that this result may be transfered to the case of invisible queues, by considering the behaviour during a single cycle from this periodical affect. Thus, without further knowledge about the queue's state, customers display decline in positive feelings. The negative feelings, described above, can be related as the price customers pay while waiting in a queue, and an increase of negative feelings in delay can be refered as an increasing cost function. We now continue to a more detailed description of the increasing behavior of the cost function.

Larson, in his article [17], describes some preliminary results of interviews that were conducted in order to assess people's disutility functions for waiting time in queues. The results showed nonlinear dependence of the disutility function on delay (9 of 10 interviews). Diekmann and Jungbauer-Gans checked drivers' reaction to a blocking car [4]. The estimated survival functions of response time showed a low rate of reaction for the first seconds, and then a steep change, which means loss of patience. Hueter and Swart, in their work for Taco Bell [13], studied how customers perceive the time in queue and found out that the first 5 minutes of waiting are perceived as a couple of minutes. After 5 minutes an elbow

appears in the graph and perceived time changes nonlinearly with absolute time. Palm [21] suggested a two parameters model for the irritation function of waiting customers. The irritation is increasing in time. Hence, the inconvenience function, which is an integration over the irritation, is convex (t power constant). The accumulated cost function, that arises from all the above, increases slowly at the beginning, but at some point changes into steep increase.

Having suggested the shape of the function, we shall now try to understand why and when the steep change (elbow) occurs. In [27] we see a distinction between two groups: high goal and low goal oriented groups, we will return to this point later on, and now examine the results for the low goal oriented group. Two main conclusions were reached regarding this group: (1) The longer people from this group waited the less pleasant of a mood they had. (2) They tended to underestimate remaining waiting time, being far from the goal, and overestimate it in a closer position. Dube Schmitt and Leclerc in [6] showed that customers, waiting for service, grow negative feelings during the wait. Taylor showed in [25] that the stronger anger and uncertainty the customer felt the longer the time was perceived. Maister in propositions 3,4 [18], also determines the more anxiety and uncertainty a waiting customer feels the longer the time is perceived.

Let us summarize these last conclusions. Customers enter a waiting line with an objective view of time (maybe a slight tendency to optimism regarding the time they are about to wait); during the wait a process of growing anxiety, uncertainty, and other negative feeling takes place. This process has a positive feedback (the more negative feeling, the more time perceived as longer, magnifying negative feelings), which explains the nonlinearity of the impatience graph.

We are still left with the question of when does this process takeoff occurs. Maister [18] claims in his first law about service encounters, that Satisfaction equals Perception minus Expectation. This is a very intuitive law: we are satisfied whenever we perceive a higher level of service than we expected. By transforming this law to our needs, we can say that customers will be satisfied till the moment when the waiting time they perceive increases

beyond what they expected. At this point the feedback process begins. By now we have suggested the shape of the impatience graph, as slowly increasing at the beginning till an elbow leading to a much higher increase rate. The elbow appears around the expected time of service.

We shall continue by describing the difference between groups of customers. As mentioned previously, Thierry shows in his article [27] a distinction between groups of customers. His results exhibit a significant difference between customers, who have high motivation to achieve the goal (service), and customers who are less eager. The high goal oriented didn't reveal a strong connection between waiting time and mood. This group also had a more optimistic estimation about the remaining waiting time (higher hazard rate). Friedman and Friedman [7] succeeded in segmenting the customers into groups according to their willingness to pay in order to shorten the waiting time. Diekmann and Jungbauer-Gans [4] showed a partitioning due to social status, which characterized impatient behavior. Hence, we can see that different groups have different values of time. Factors like goal motivation, social status and others suspend the annoyance, and therefore extend the period of time customers would be willing to wait.

5.3.3 Properties of optimal decisions

Using the results of the abandonment rule (see subsection 5.3.1) and the shape of the cost function (see subsection 5.3.2) , we are now about to analyze two cases that are distinguished by their marginal cost function. In the first case the assumption is that customers believe in an exponential distributed waiting time with the parameter μ_z. The marginal cost function (derivative of the cost function) is an increasing function of t, varying in its rate of increase between different customers. The second case maintains the exponential waiting time assumption, but takes into consideration the influence of the customers' expectations. Here we assume that the cost function is an increasing function of $T_z = t - shft(x)$ instead of t alone. Thus, allowing a longer waiting period when the estimated waiting time increases (or a shorter one if x decreases). While we expect an increase in the waiting period as x increases, it is reasonable that the rate of increase in patience will not exceed the increase rate of x. Therefore, we add one more constraint on the shift function: $\frac{d(shft(x))}{dx} \leq 1$.

Proposition 5.5 *Given the utility function (5.14) and under the exponential waiting time belief with an increasing cost-benefit ratio function - $\gamma_z(t)$, the optimal abandonment time of a type z customer is:*

(i) $T_z = 0$ *for* $\gamma_z(0) \geq H_z$

(ii) $T_z = \infty$ *if* $\gamma_z(t) \leq H_z$ *for every* t.

(iii) $T_z = \gamma_z^{-1}(H_z)$ *if there exists* t *which satisfies* $\gamma_z(t) > H_z$.

Proof: As shown in (5.18) $H_z = \mu_w$. We are looking for the maximum value of t, which maintains $\gamma_z(t) \leq H_z$.
If $\gamma_z(0) \geq H_z$, and knowing that $\gamma_z(t)$ is increasing, the utility function would be decreasing (negative derivative). Therefore, in order to maximize the utility we must abandon at $T_z = 0$.

If $\gamma_z(t) < H_z$ for every t the Utility function is an increasing function. Therefore, the optimal abandonment time is $T_z = \infty$.
If there exists t which satisfies $\gamma_z(t) > H_z$ the utility function is an increasing decreasing function. The derivative is positive till $\gamma_z(t) = H_z$ and then becomes negative. Hence, the optimal abandonment time would be $T_z = \gamma_z^{-1}(H_z)$. □

As previously argued, customers enter the system with expectations that we refer to as the estimated mean waiting time. The last cost function does not take into consideration those expectations, because it depends on t alone. Hence, we suggest (see section (5.2)) the next cost function, which shifts the nonlinear increase of the cost function relative to the expectations. Two assumptions appear in this set of functions: the shift behavior, and its bounded increase rate. Both assumption are mainly intuitive. The shift function is a simple way to express the increase in patience as a result of increase in the estimated waiting time, while bounding the increase rate is only reasonable since we do not expect customers to increase their relative willingness to wait.

Proposition 5.6 *Let $\gamma_z(t) > 0$ be an increasing differentiable cost-benefit ratio function, and let $T(x)$ be a function which satisfies the equation $\gamma_z(T(x) - x) = 1/x$. Then the increase rate of $T(x)$ is bounded by 1. In particular for $\gamma_z(T(x) - shft(x)) = 1/x$ where $\frac{d(shft(x))}{dx} < 1$, where $\frac{\partial T}{\partial x} < 1$.*

Proof: Using the inverse function γ_z^{-1}, the direct expression for $T(x)$ is:

$$T = x + \gamma_z^{-1}\left(\frac{1}{x}\right)$$

Differentiating $T(x)$ with respect to x:

$$\frac{\partial T}{\partial x} = 1 - \frac{1}{x^2}\frac{\partial \gamma_z^{-1}\left(\frac{1}{x}\right)}{\partial\left(\frac{1}{x}\right)}$$

Since $\gamma_z(t) > 0$ is an increasing function in t, $\gamma_z^{-1}(t)$ is also increasing. Therefore, the derivative of γ_z^{-1} is positive, and so $\frac{\partial T}{\partial x} \leq 1$. It is clear that if instead of x we use $shft(x)$ with $\frac{d(shft(x))}{dx} < 1$, we get $\frac{\partial T}{\partial x} < 1$. □

After having described two types of abandonment time functions, we continue discussing the patience function behavior. The equilibrium theorems will be based on the attributes, which are derived through the relation between the individual abandonment time and the global patience function.

6 Equilibrium Uniqueness

We now turn to analyze the equilibrium. In the model formulation section we have suggested one parameter, to describe the statistics of the subjective waiting time (instead of the distribution function in the consistent equilibrium).We have also claimed that since x reflects only the knowledge available to individuals rather than complete understanding of the system, it may have a different value than the calculated mean in the system. Therefore, we defined the partial consistency assumption, as the state where the estimated and calculated waiting time are identical.
In the patience function section we have explained that it is more natural to refer to the conditional estimated mean waiting time - $E_x[v|v > 0]$, since there is no reason to address an hypothetical patience function, if a customer does not need to wait for service. Under this assumptions the equilibrium we are about to refer to is the *Partial Consistent Equilibrium.*

Partial Consistent Equilibrium

Definition 2 *The system is in a partial consistent equilibrium, if the estimated and conditional calculated waiting time coincide for all customers:*

$$E_x[v|v > 0] = x$$

The following theorems describe the conditions for equilibrium. First, the relation between the patience function and the estimated waiting time is explored. Then, the relation found is used to show the uniqueness of the equilibrium for the rational decision rule.

Theorem 6.1 *Consider an $M/M/m + G$ queue with $\bar{G}(x,t)$ as a decreasing function with x. Then there exists a unique partial consistent equilibrium.*

Proof: We have proven in Proposition (5.1) that a decreasing function $\bar{G}(x,t)$ in x results in a decreasing calculated mean function: $E_x[v|v > 0]$. It is easy to see that there is a

unique x, which satisfies $E_x[v|v>0]=x$. But this is the partial consistent equilibrium. □

Theorem 6.2 *Let $\bar{G}(t-shft(x))$ be the patience function in an M/M/m+G queue. Assume also that $shft(x)$ is non-decreasing and with $\frac{d(shft(x))}{dx}<1$. Then there exists a partial consistent equilibrium, it is unique.*

Proof: In proposition (5.4), we have stated that $E_x[v|v>0]$ is increasing in x, with a less than one increase rate. A partial consistent equilibrium satisfies $E_x[v|v>0]=x$. The increase rate of x in x is 1, and knowing that the increase rate of $E_x[v|v>0]$ is less than 1, guarantees that $\exists x_0>0$ which satisfies $E_{x_0}[v|v>0]=x_0$, and that there is no other x, which satisfies the equilibrium equation.

□

Uniqueness of the equilibrium under the rational decision rule can now be established. We start with a marginal cost-benefit ratio function of t alone:

Theorem 6.3 *Let $\gamma_z(t)$ be a time dependent increasing cost-benefit ratio function, and assume that customers believe in an exponential waiting time distribution. Then there exists a unique partial consistent equilibrium.*

Proof: Proposition 5.5 gives rise to the value of T_z. The cost-benefit ratio function is known to be increasing, therefore T_z is increasing with $1/x$, or decreasing with x. Decrease of T_z with x means decrease of $\bar{G}(x,t)$. In theorem 6.1 we have proved that if $\bar{G}(x,t)$ is a decreasing function with x, therefore there exists a unique partial consistent equilibrium.

□

We have proposed a model, which shifts the time dependent marginal cost-benefit ratio function, according to x. The following theorem provides conditions for uniqueness of the equilibrium.

Theorem 6.4 *Consider the M/M/m queue with the rational abandonment model, where $\gamma_z(T-shft(x))>0$ is the increasing cost-benefit ratio function, with $\frac{d(shft(x))}{dx}<1$. If there exists a partial consistent equilibrium it is unique.*

Proof: As shown in proposition 5.6 an increasing $\gamma_z(T - shft(x))$ leads to the bounded derivative: $\frac{\partial T}{\partial x} < 1$. Equivalently we can say that the shift of the patience function is bounded by x. According to theorem 6.2, this means that there exists a unique equilibrium.

□

The last theorem proved uniqueness, but if we think about human customers an infinite willingness to wait in the queue is inconceivable. Therefore, it is required to add some threshold, after which the customers will not agree to wait any longer. Adding such threshold insures the existence of the equilibrium according to the remark in theorem 6.2.

7 Survival Analysis

Survival analysis is concerned with the measurement of time between an initial event and an end event - a process lifetime. This period time is known as the "survival time". The name was chosen since the technique emerged out of the insurance industry, where risk calculations were needed for costing the insurance premium [22]. Survival data has an important feature - censored measurements, as explained below. Together with the ordinary data of the length of the interval, some of the information is partial, due to ceasing the process before the end event.

It is important to note that the survival analysis can be used either to estimate the offered waiting time, in which case the censored samples are the abandonment samples. It can also be used as for estimating customers patience, in which case the samples terminated by service are the censored ones. In our simulation we estimate the offered waiting time, therefore the first case applies.

7.1 Censoring in waiting time estimation

When we come to estimate the mean waiting time in a queue with abandonment, we must consider two types of waiting experiences. One would be a waiting period ended by admission into service, while the other - a waiting period terminated by abandonment. The abandonment time, although not specifying the exact waiting duration till admission, supplies us with a lower bound to this value. Samples like the abandonment times, which reveal only partial information, are referred to as censored samples.

It is clear that we can not relate to the censored samples the same way as regular samples, neither can we ignore them. abandonment occur when the waiting periods are too long for the customers patience. Therefore it is most probable that instead of high values of complete waiting periods we will get lower values of abandonment times. Ignoring the censored samples or referring them as regular samples will result in a biased estimator. In

the next subsections we describe two estimators, which relate to censored sample as well: Kaplan-Meier estimator, and censored maximum likelihood estimator.

7.2 Kaplan-Meier Estimator

The Kaplan-Meier Estimator enables us to find an estimation for the waiting time distribution, based on censored data. Denote $\hat{S}(t)$ as the estimator of $\bar{F}(t) = P(V > t)$, where V is the waiting duration till admission to service. Let us also divide the time scale to finite intervals of size Δ, and let time t_i equal $i\Delta$. We can now write $\hat{S}(t)$ as follows:

$$\hat{S}(t_i) = P(V > t_i | V > t_{i-1}) P(V > t_{i-1})$$

The same way we can extend the last formula for every $t_j < t_i$:

$$\hat{S}(t_i) = \prod_{j=1}^{i} P(V > t_j \,|V > t_{j-1})$$

The information we have regarding the waiting experiences of customers is the samples of completed waiting durations or the abandonment times. The conditional probability of a waiting duration V to be longer than t_i, given that it is longer than t_{i-1} is:

$$P(V > t_j \,|V > t_{j-1}) = \frac{n_j - h_j}{n_j}$$

where:

n_j - the number of trials, which were neither completed nor censored before t_j.

h_j - the number of trials, which were completed by the time t_j.

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{n_j - h_j}{n_j}$$

As mentioned before, we are using the estimation in order to find the mean waiting time. Using the Kaplan-Meier estimator, we first estimate the waiting time distribution, and then calculate the mean waiting time through the tail formula:

$$E[v] = \int_0^\infty \bar{F}(v) dv$$

7.3 Censored MLE Estimator

The maximum likelihood function in the censored case takes the following form:

$$CMLE = \max_{\mu} \left\{ \prod_{i=1}^{m} [f_\mu(x_i)\Delta_i] \prod_{j=1}^{n-m} [1 - F_\mu(T_j)] \right\}$$

The notations are:

n - the total number of samples.

m - the number of complete waiting durations.

x_i - complete waiting duration sample.

T_i - abndonment time sample.

Δ_i - small time interval around x_i.

Since the time intervals do not depend on the parameter - μ, we can rewrite the likelihood function:

$$CMLE = \max_{\mu} \left\{ \prod_{i=1}^{m} [f_\mu(x_i)] \prod_{j=1}^{n-m} [1 - F_\mu(T_j)] \right\}$$

Assuming an exponential distribution with the parameter - μ:

$$f_\mu(x_i) = \frac{1}{\mu} \exp(-\frac{x_i}{\mu});\ F_\mu(T_i) = 1 - \exp(-\frac{T_i}{\mu}),$$

and the censored likelihood function:

$$CMLE = \frac{1}{\mu^m} \exp\left(\frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n-m} T_j}{\mu} \right)$$

Differntiating and comparing to zero, we find that the censored maximum likelihood function yields the following estimator for the offered waiting time:

$$\hat{\mu} = \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n-m} T_j}{m}$$

8 A learning model - formulation and simulation

We suggest a simple learning model based on our behavioral assumption - the estimated mean waiting time as the main factor influencing customers' patience. According to the model, customers learn the queue statistics through their experience of entering service or abandoning the queue. We use simulation to help us, even partially, overcome the problem of obtaining the detailed data needed to examine the theoretical model. Through simulation, we can check the behavior of the system under different parameters, compare customers decision rules and mean estimation techniques etc. We can see whether the system converges, and check for correlation between the converged-to values and the mathematical results.

The simulation is of an M/M/1 + G queue with hetergeneous customer population. Customers arrive to a queue (infinite size) at times that are determined according to a Poisson process with parameter λ. The service in the system is FIFO, and its duration is exponentialy distributed with the parameter $1/\mu$. The patience function of every customer is translated into a time T, after which the customer will abandon the queue. Customer admitted into service can no longer abandon the system. The type of a customer is uniformly selected from the set of allowed types. Given the type of customer, the customer is uniformly selected from the bank of this type group of customers. If all the customers from this group happen to be in the queue, a new customer is initialized with the knowledge-base of a uniformly chosen existing customer. Every new type customer begins the queue experience with ten initial offered waiting periods set to be T_0. This is in order to reduce the number of iterations needed before the system converges. We also allow a periodic exploration of the waiting time by substituting the real abandonment time with 1/eps; this accelerates learning.

The abandonment rule is determined as a linear function of the estimated waiting time with a random element: $T(x) = \alpha * x + U$. Here U is a Normally distributed random number with mean zero and variance 0.01. The estimated waiting time is calculated with

two possible estimators: Censored MLE or Kaplan-Meier.

8.1 Simulation results

The following results consist of four examples. The first two examples compare system behavior under two different individual mean waiting time estimators, with a single type of customers (given the same queue experience all customers would choose the same abandonment time). We show convergence of both systems, but we also show that the Kaplan-Meier estimation result is closer to the theoretical equilibrium value. The next two examples, again, compare the two estimators, but in a system with two types of customers. We show that there is a correlation between the estimated waiting time value of the customers and their customer type. It is important to note that we do not prove convergence mathematically, although dozens of simulations yielded the same convergence results.

Example 1 The parameters of the following simulation results are: $\lambda = 1$, $\mu = 1$, $T_0 = 1.5$ (initial 10 values of trials in the system), $Z = 1$ (one type of customers), exploration period $= 30$, $T(x) = 0.8 * x$, Kaplan-Meier estimator.

In the next pages we present the estimation of the waiting time distribution, as was performed by every customer. We also show the convergence of the estimated waiting time through every trial a customer had in the system (iteration).

In figures 1,2 we can see the estimated waitining time distribution of each customer:

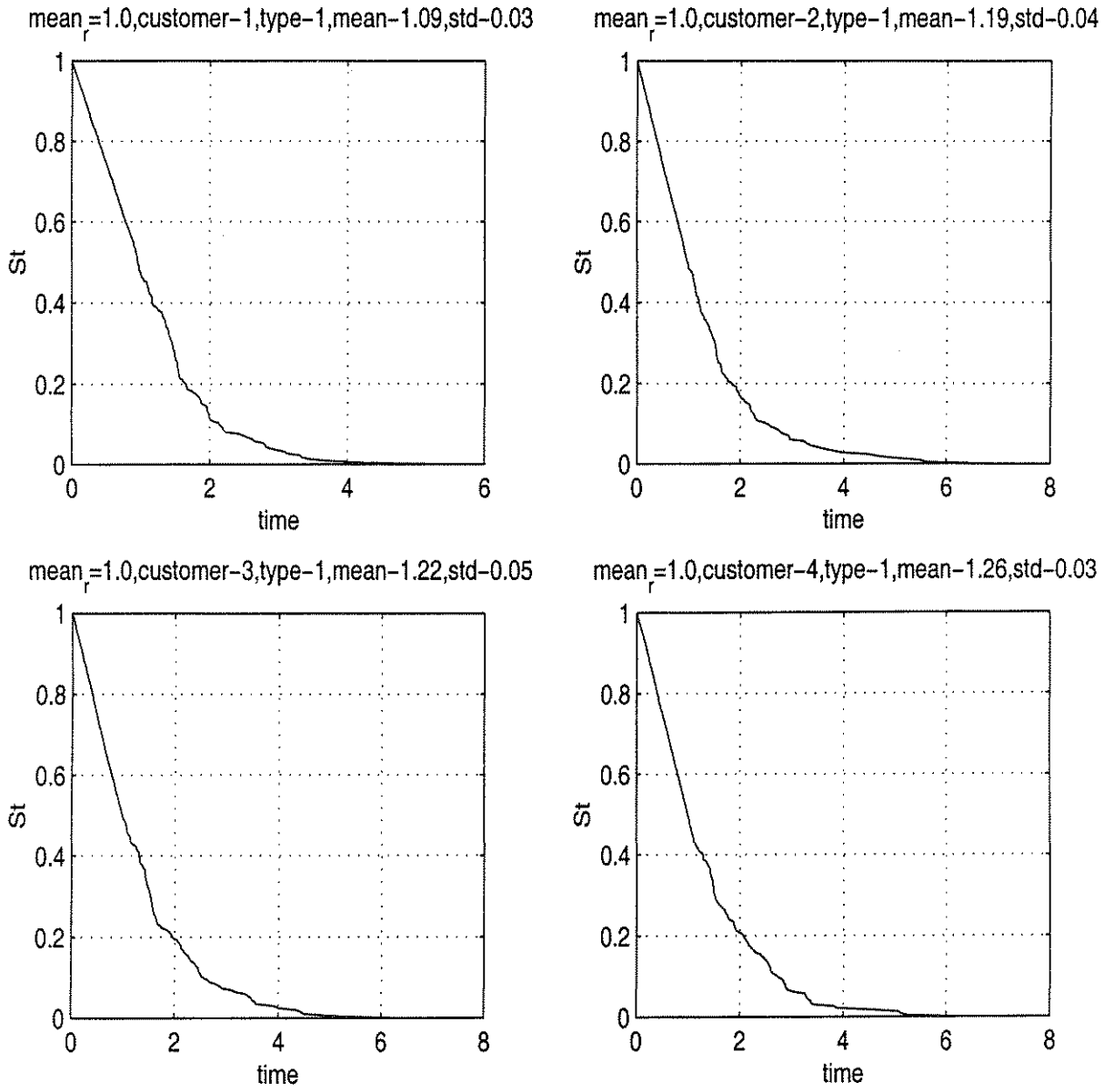

Figure 1: *estimation of the waiting time distribution customers 1-4*

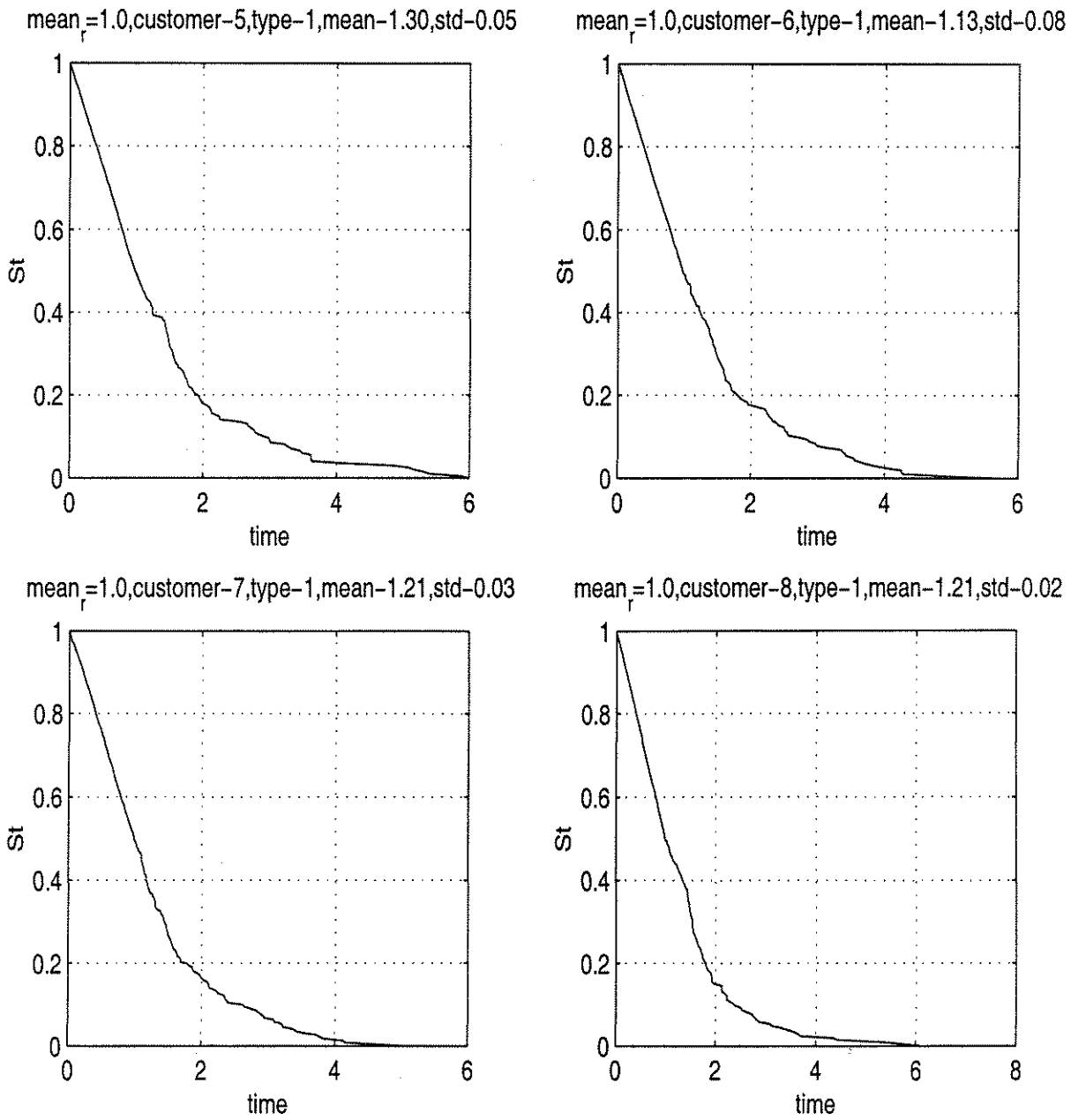

Figure 2: *estimation of the waiting time distribution customers 5-8*

Figures 3,4 show the estimated mean waiting time of the 8 customers:

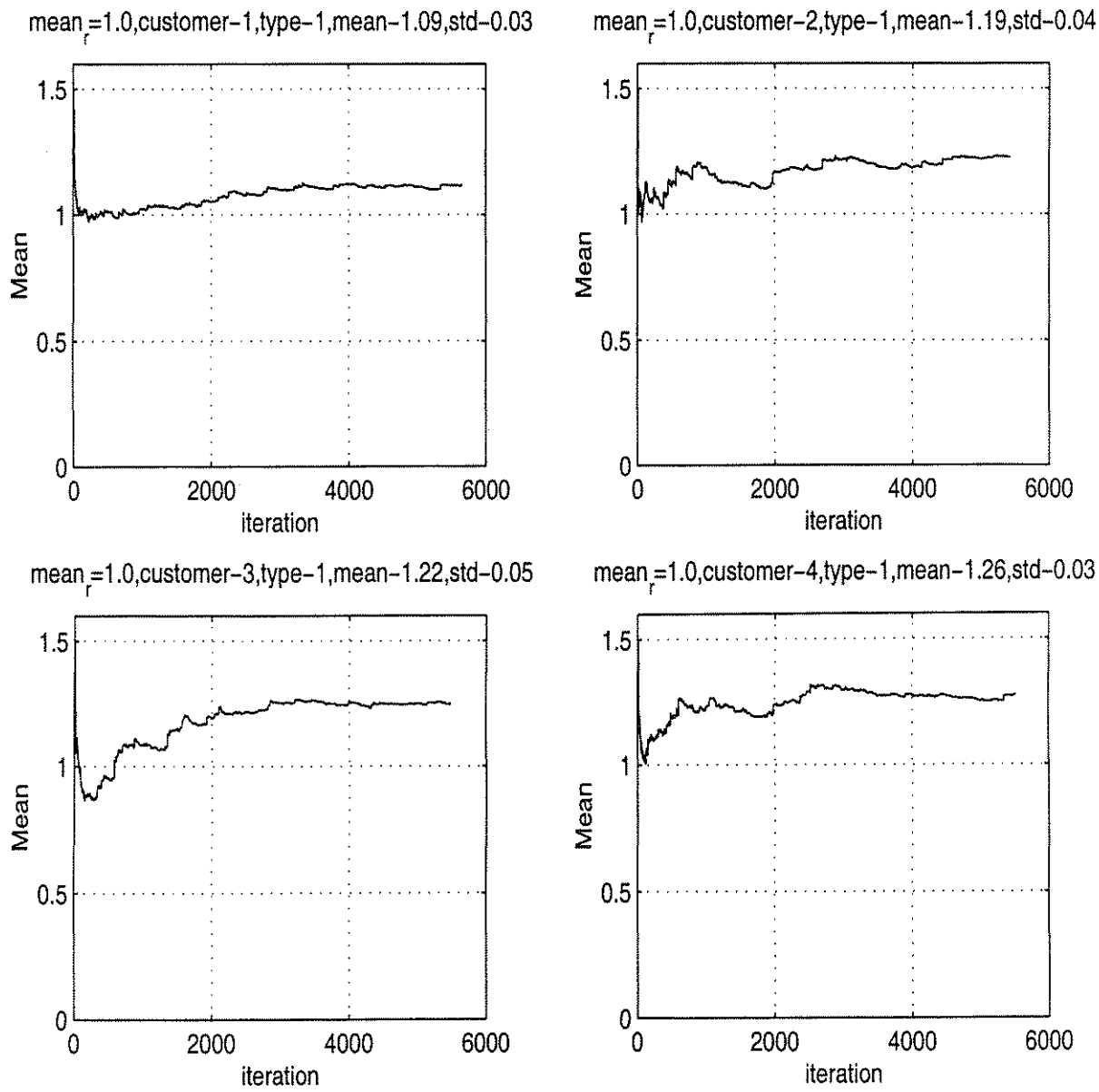

Figure 3: *estimation of the waiting time customers 1-4*

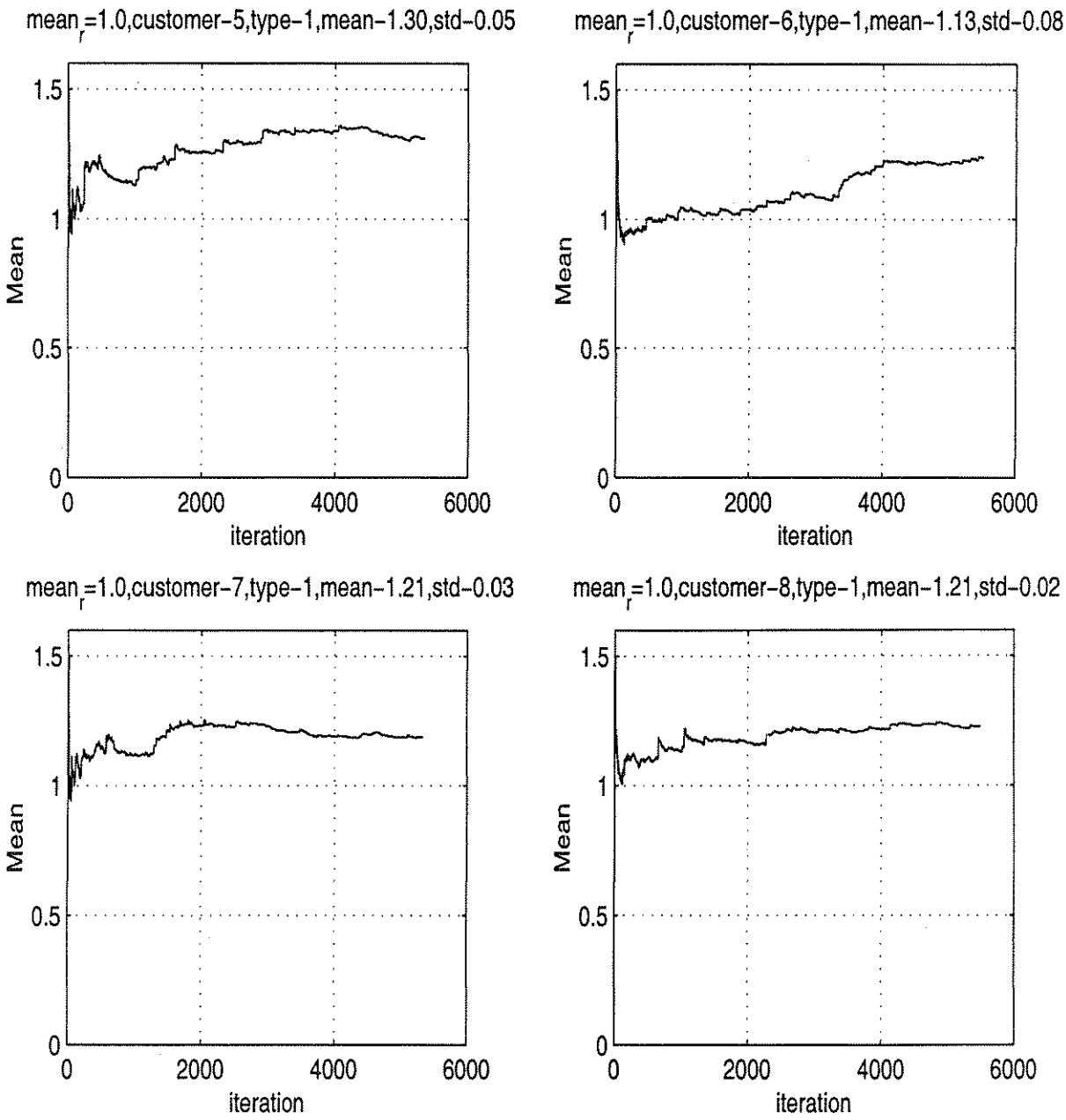

Figure 4: *estimation of the waiting time customers 5-8*

We see that the estimated waiting time converges. The simulation yields a mean waiting time of 1.2007 across 8 customers with standard deviation of 0.0672.

The theoretic conditional calculated waiting time for one type of customers uses:

$$\bar{G}(x,t) = \begin{cases} 1 \text{ for } t \leq T \\ 0 \text{ for } t \geq T \end{cases} \tag{8.19}$$

we get:

$$E_x[v|v>0] = \frac{\int_0^\infty t \exp(-\int_0^t \mu - \lambda \bar{G}(x,s)ds)dt}{\int_0^\infty \exp(-\int_0^t \mu - \lambda \bar{G}(x,s)ds)dt} = \frac{\int_0^T t \exp\left((\lambda-\mu)t\right)dt + \exp\left(\lambda T\right)\int_T^\infty t\exp(-\mu t)dt}{\int_0^T \exp\left((\lambda-\mu)t\right)dt + \exp\left(\lambda T\right)\int_T^\infty \exp(-\mu t)dt}$$

Calculating the integrals:

$$E_x[v|v>0] = \frac{\frac{\lambda}{\mu}T\left(x\right)\exp\left(cT\left(x\right)\right) + \frac{a}{2c}\exp\left(cT\left(x\right)\right) + \frac{1}{c}}{\frac{\lambda}{\mu}\exp\left(cT\left(x\right)\right) - 1}$$

where

$$a = \frac{2\lambda(\lambda-2\mu)}{\mu^2};\ c = \lambda - \mu;$$

For the case of $\lambda = \mu = 1$:

$$E_x[v|v>0]\,|_{\lambda=\mu=1} = \frac{\frac{T(x)^2}{2} + T\left(x\right) + 1}{T\left(x\right) + 1}$$

In our simulation we used the linear abandonment time: $T\left(x\right) = 0.8x$. Therefore in order to find the theoretic equilibrium $E_x[v|v>0] = x$, we need to solve the following equation:

$$\frac{\frac{(0.8x)^2}{2} + 0.8x + 1}{0.8x + 1} = x$$

The solution is a mean waiting time of $x = 1.25$, which is indeed close to the simulation's result.

The Kaplan Meier estimator seems to bring the customers near the unique partial consistent equilibrium, but is it reasonable to assume that customers use such a complex estimator? In the next example we use a simpler more intuitive estimator - Censored MLE estimator. For this estimator we first assume an exponential distribution (instead of trying to find the correct one), and then calculate the estimator of the parameter (mean) using the MLE.

Example 2 The parameters are the same as in the last example ($\lambda = 1$, $\mu = 1$, $T_0 = 1.5$, $Z = 1$, exploration period $= 30$, $T(x) = 0.8 * x$) except for the Censored MLE estimator.

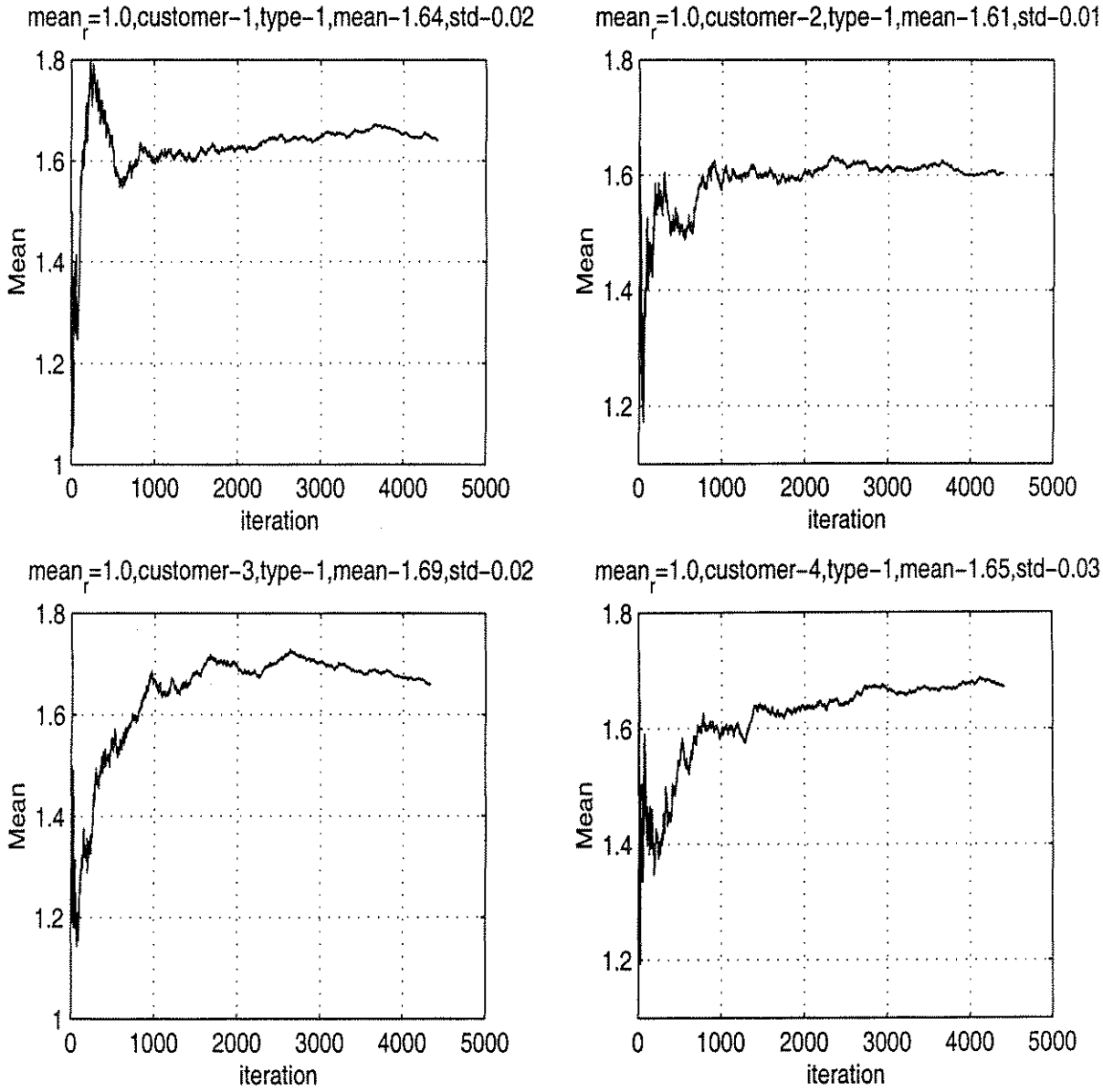

Figure 5: *estimation of the waiting time 1-4*

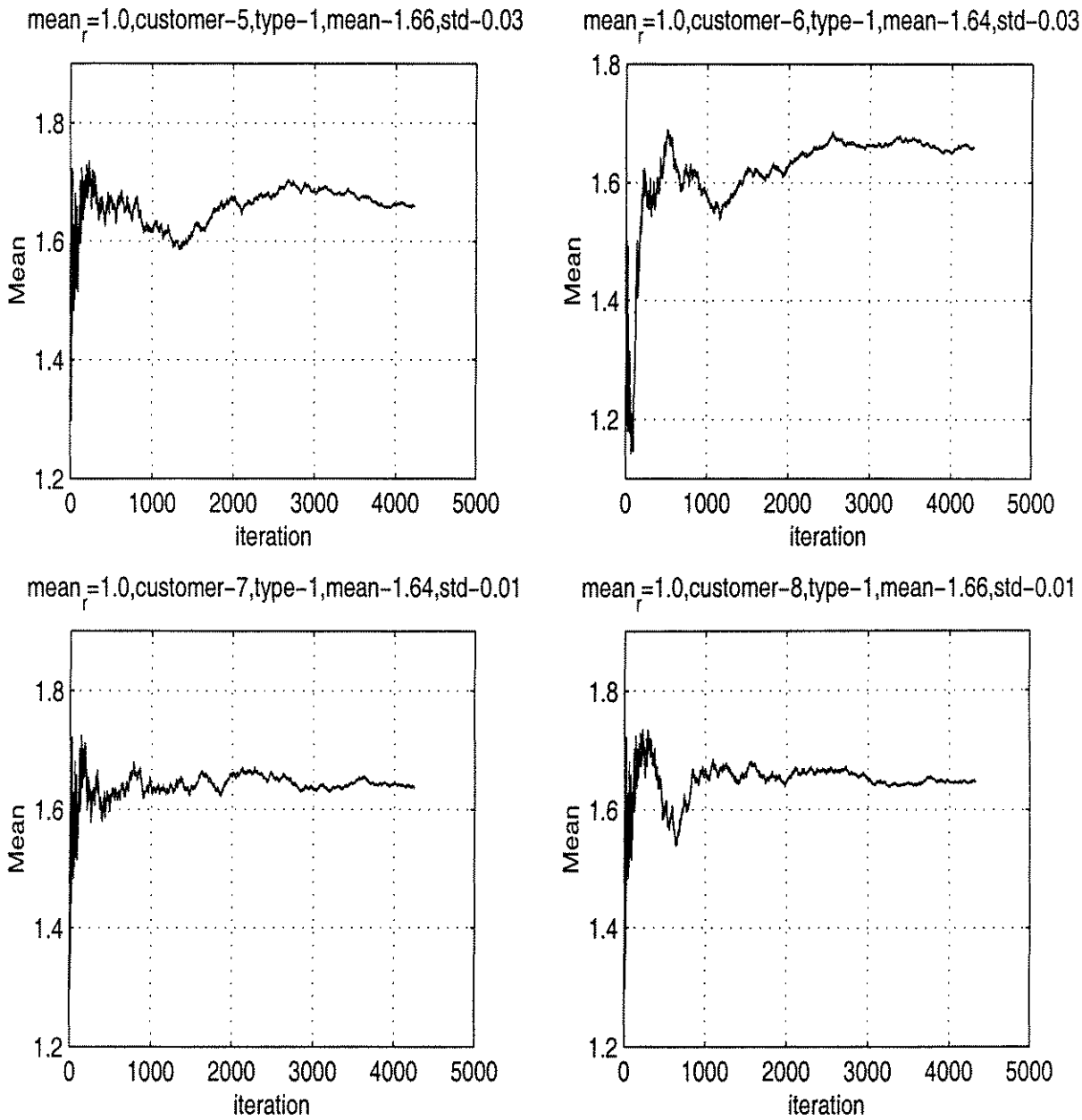

Figure 6: *estimation of the waiting time 5-8*

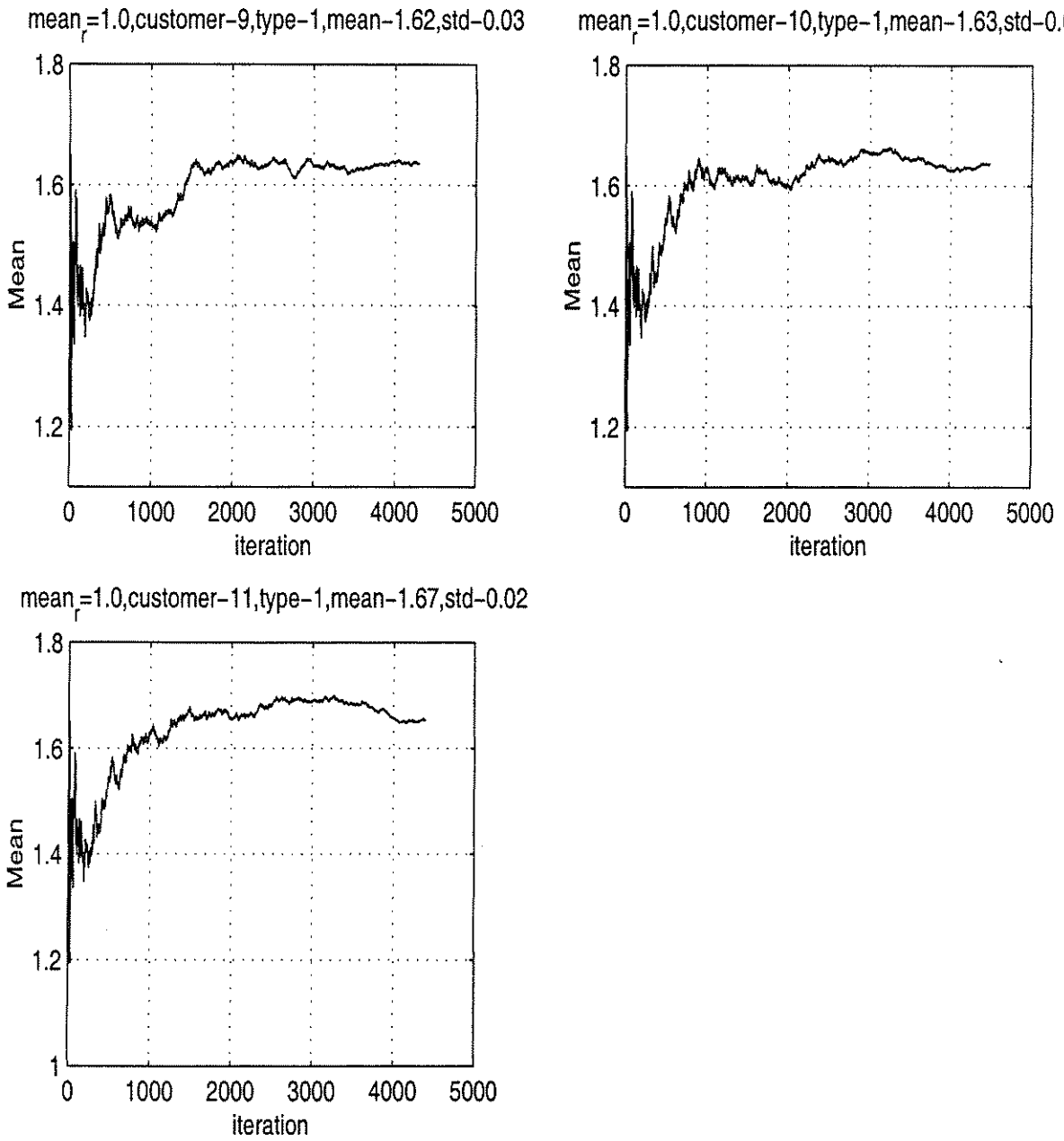

Figure 7: *estimation of the waiting time 9-11*

We see that the estimated waiting time converges. The simulation yields a much higher mean waiting time of 1.6452 across 11 customers, with standard deviation of 0.0218. This result is not surprising, since the base assumption of the estimator was not correct - the waiting time distribution is not exponential because of the abandonment. Recall the shape of the distribution, as was estimated using the Kaplan-Meier estimator in example 1. The next figure shows the a-parametric estimation of the waiting time distribution function, with a comparison to an exponential distribution with the same mean (the results are for customer number one. The other customers showed similar results regarding the distribution, as seen in figures 1,2). The smooth curve is the curve of the exponential distribution:

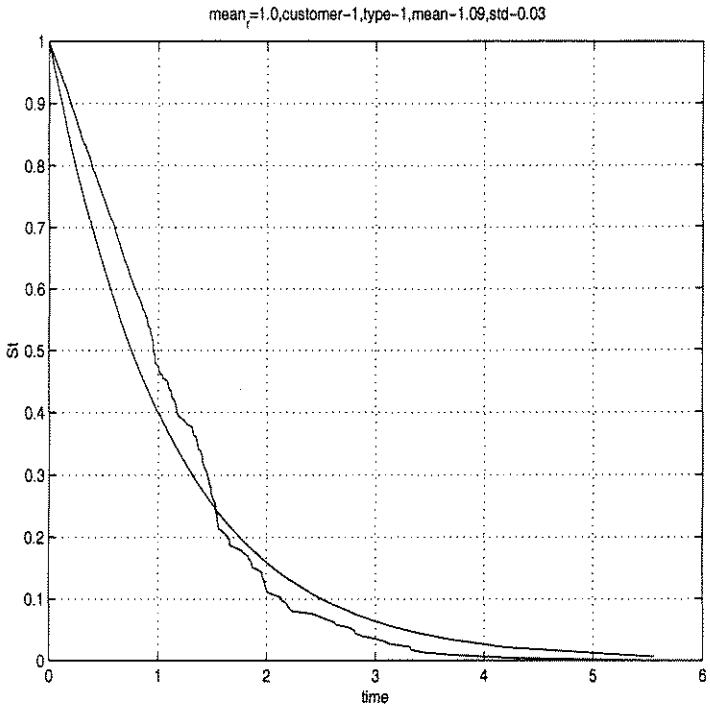

Figure 8: *estimation of the waiting time distribution*

Notice that for $t < 1.5$ the "true" distribution in figure 8 has higher values than the exponential distribution. The mean estimated waiting time for the censored MLE is 1.6452, which means abandonment time of $T = 0.8 * 1.6452 = 1.31616$. Therefore, the full waiting time durations (the durations which ended with service, not abandonment) reflect only the part of the distribution for which $t < 1.31616$. The outcome is a higher than exponential distribution curve - a larger parameter.

The theoretic calculation of the equilibrium value is based on the ratio of abandonment and service samples. The estimated mean waiting time x using the censored MLE is:

$$x = \frac{\sum_{i=1}^{m} x_i}{m} + \frac{\sum_{j=1}^{n-m} T_j}{m}.$$

If T is the mean abadonment time, $\bar{G}(x,t)$ from (8.19) and m is large enough, then $\frac{\sum_{i=1}^{m} x_i}{m}$ equals the mean waiting time in the interval $(0,T]$, which is:

$$\left.\frac{\int_0^T t \exp(-\int_0^t \mu - \lambda\bar{G}(x,s)ds)dt}{\int_0^T \exp(-\int_0^t \mu - \lambda\bar{G}(x,s)ds)dt}\right|_{\mu=\lambda=1} = \frac{\int_0^T tdt}{\int_0^T dt} = \frac{T}{2}$$

We know that $T = 0.8 * x$, therefore:

$$x = \frac{\sum_{i=1}^{m} x_i}{m} + \frac{\sum_{j=1}^{n-m} T_j}{m} = \frac{T}{2} + \frac{(n-m)T}{m} = 0.4x + \frac{0.8(n-m)x}{m}$$

The last equation yields a ratio of 3/4 between the number of abandonment and the number of services.

Another way to calculate this ratio is through the waiting time distribution:

$$\frac{P(t > T)}{P(t < T)} = \frac{\exp(T)\int_T^\infty \exp(-s)ds}{\int_0^T ds} = \frac{1}{T}$$

And we get $T = 1.333$ or $x = 1.666$, which is very close to the simulation result: 1.6452.

The x here stands for the estimated mean waiting time, and not the correct mean in the system, since, as we said before, the exponential assumption is not true.

The first two examples considered only one type of customers. The next two use the same parameters, but add another type of customers. Customer of type 1 now has the patience according to $T(x) = 0.95 * x$, and customer of type 2 has $T(x) = 0.65 * x$. The results are shown in figures 10-18 for Kaplan-Meier and 19-22 for Censored MLE.

Example 3 The parameters are: $\lambda = 1$, $\mu = 1$, $T_0 = 1.5$, $Z = 1$, exploration period $= 30$, $T_1(x) = 0.95 * x$, $T_2(x) = 0.65 * x$ Kaplan-Meier estimator.

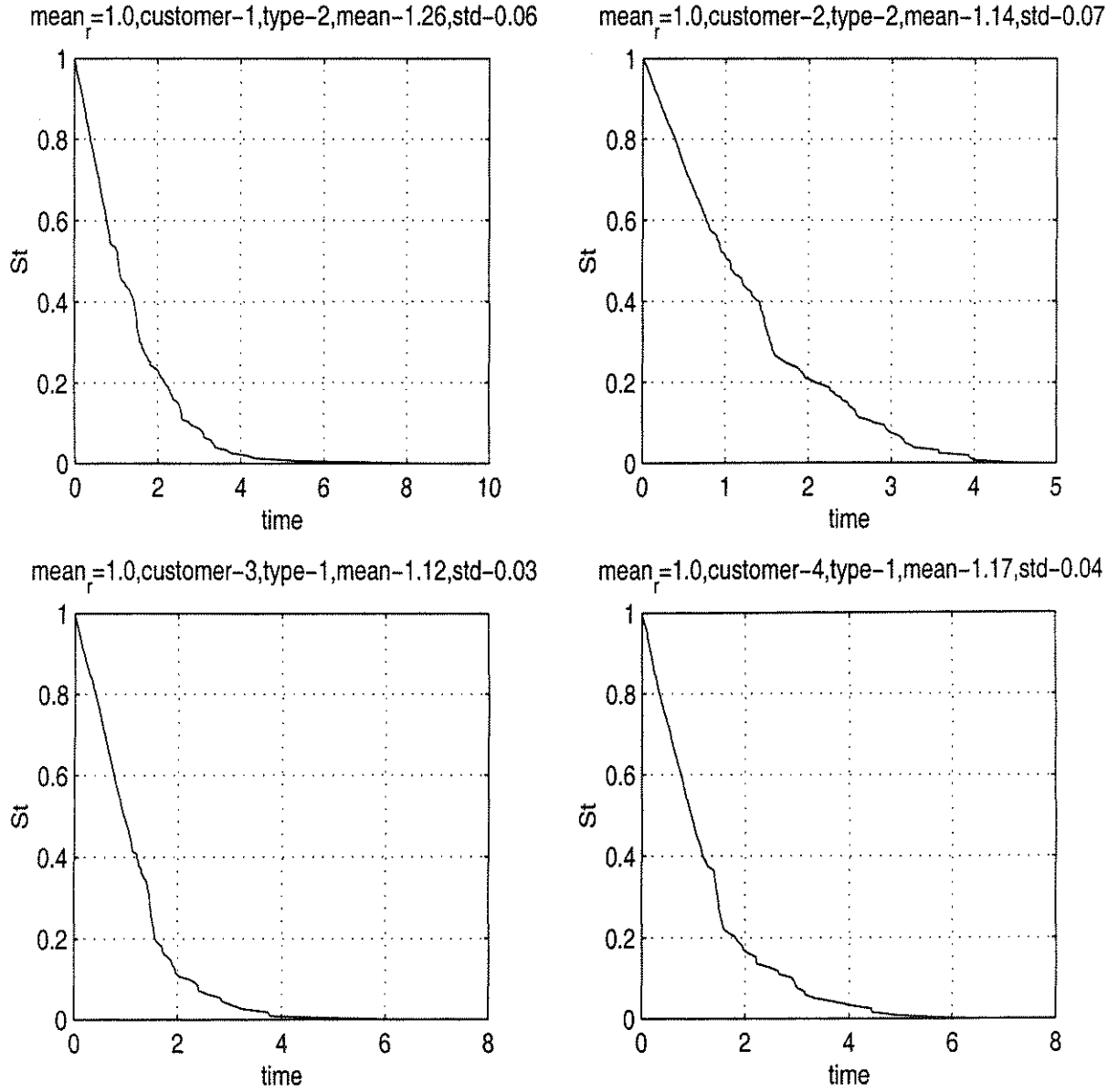

Figure 9: *Kaplan-Meier estimation of the waiting time distribution, customers 1-4*

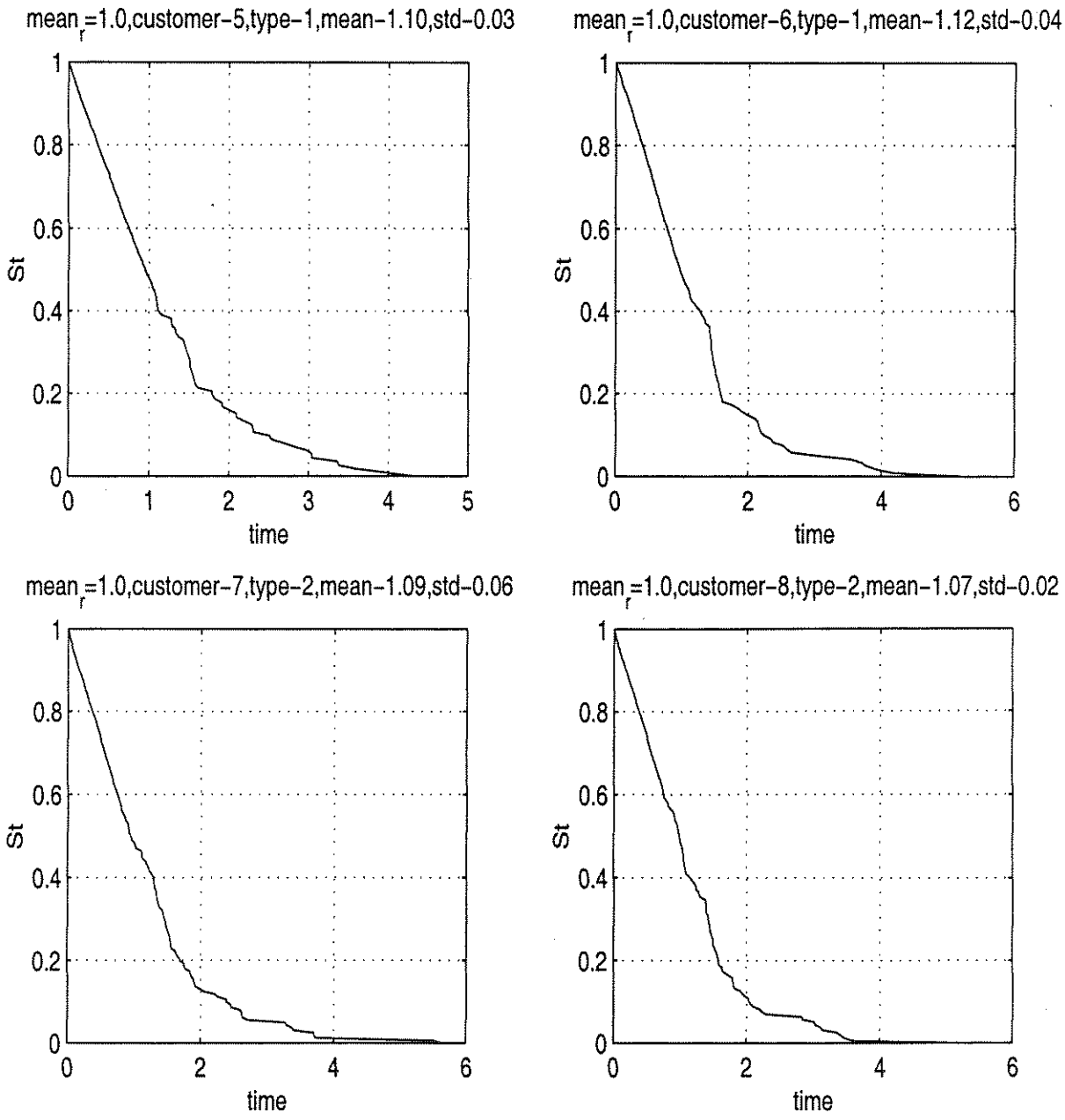

Figure 10: *Kaplan-Meier estimation of the waiting time distribution, customers 5-8*

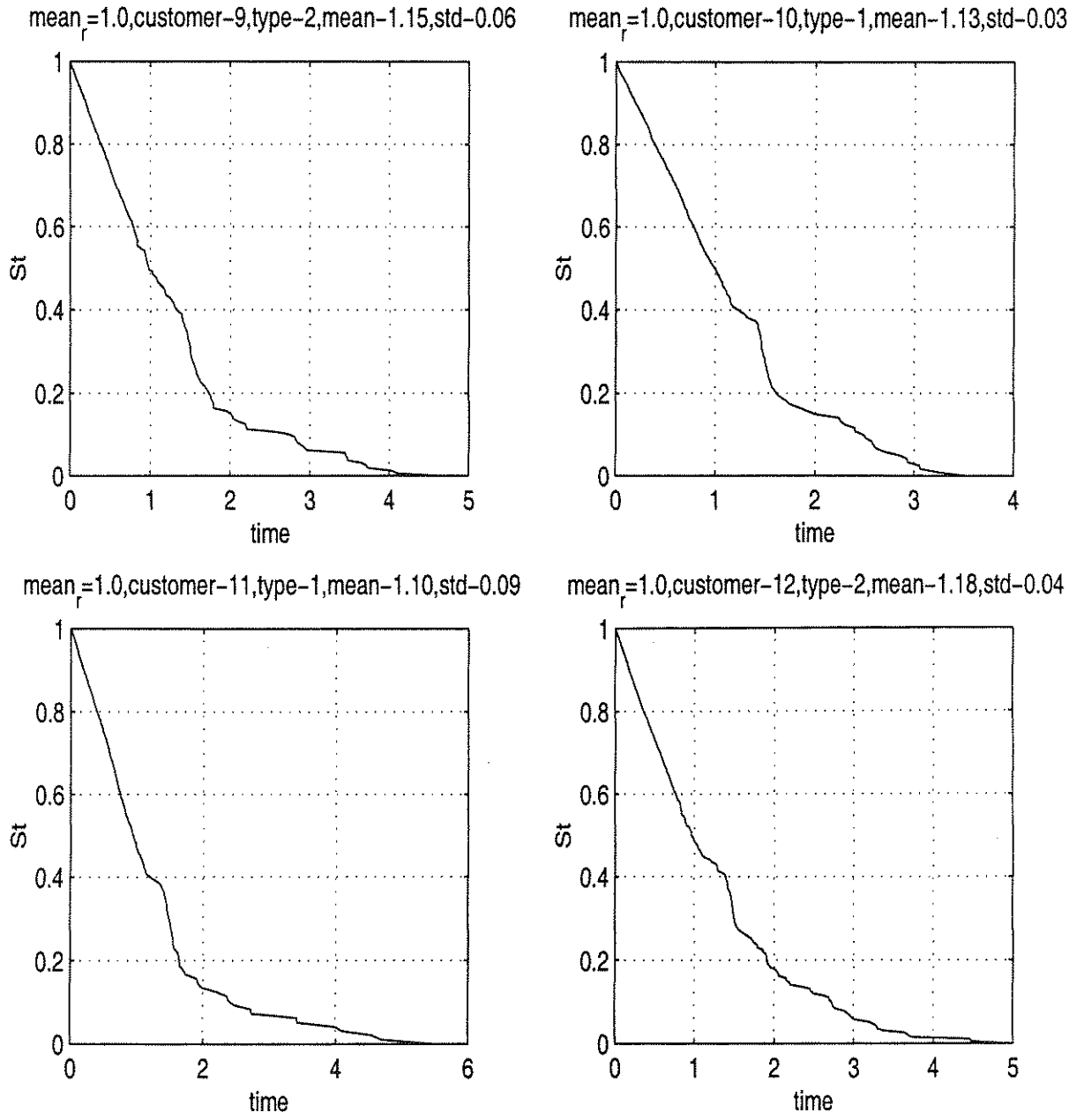

Figure 11: *Kaplan-Meier estimation of the waiting time distribution, customers 9-12*

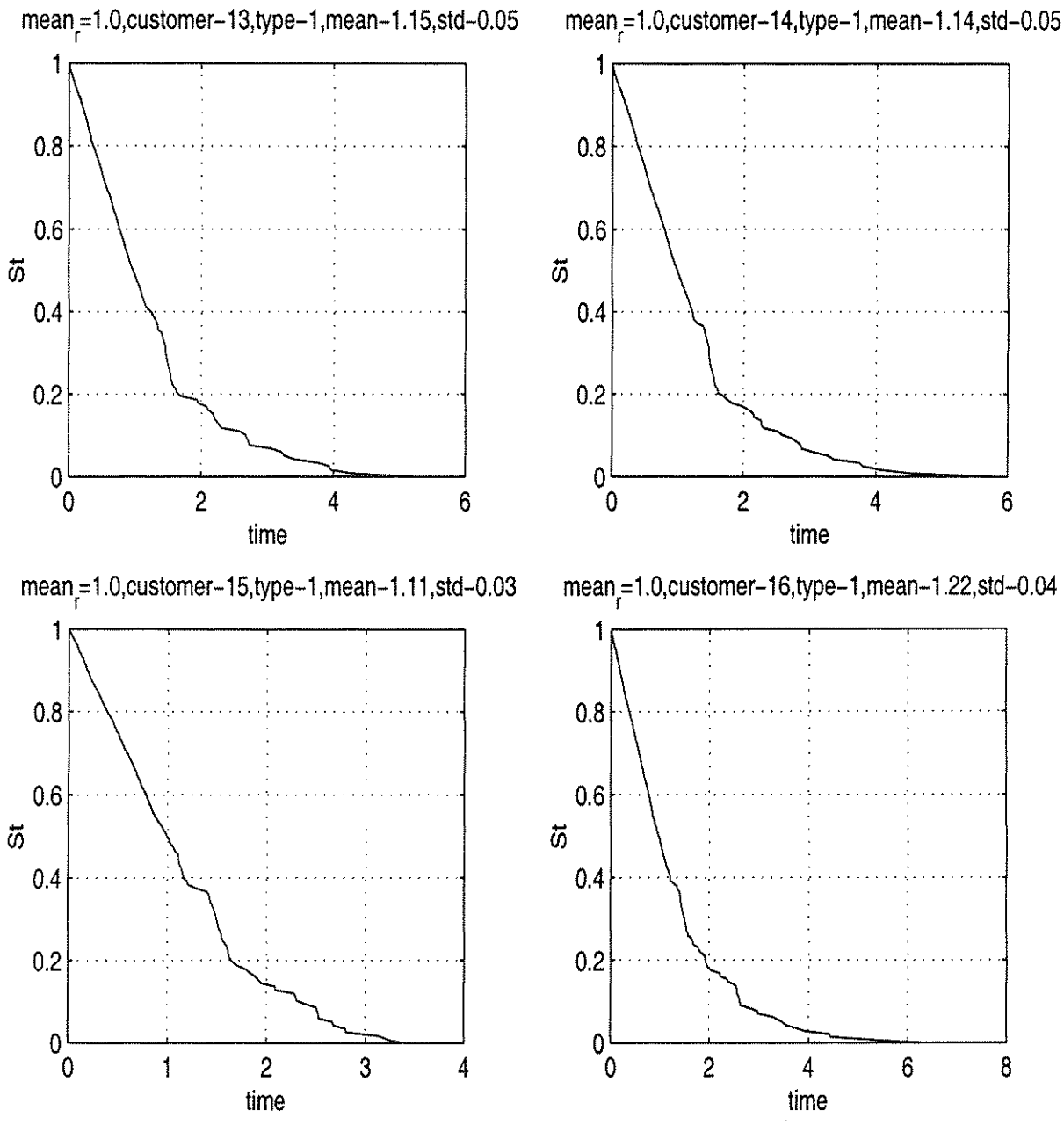

Figure 12: *Kaplan-Meier estimation of the waiting time distribution, customers 13-16*

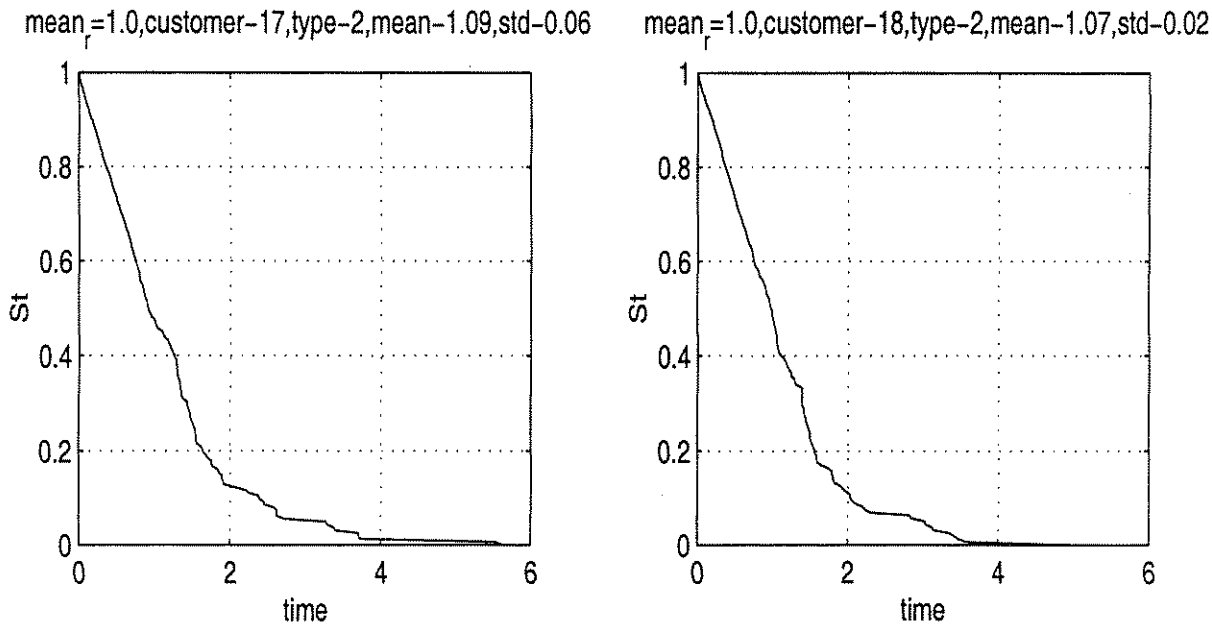

Figure 13: *Kaplan-Meier estimation of the waiting time distribution, customers 17-18*

The mean waiting time of all 14 customers is 1.1335 with standard deviation of 0.0491. The mean waiting time of type 1 customers is 1.1350 with standard deviation of 0.0368. The mean waiting time of type 2 customers is 1.1316 with standard deviation of 0.0641. We see that the Kaplan-Meier estimator yields similiar results for both types of customers. As explained before this estimator does not assume the incorrect exponential assumption, but performs estimation to the distribution itself. If the estimation is correct and the system converges, then all customers should reach the same distribution and therefore the same mean estimated waiting time, although each type acts in a different way under the same mean: type 1 abandons after 1.078, while type 2 after 0.7356.

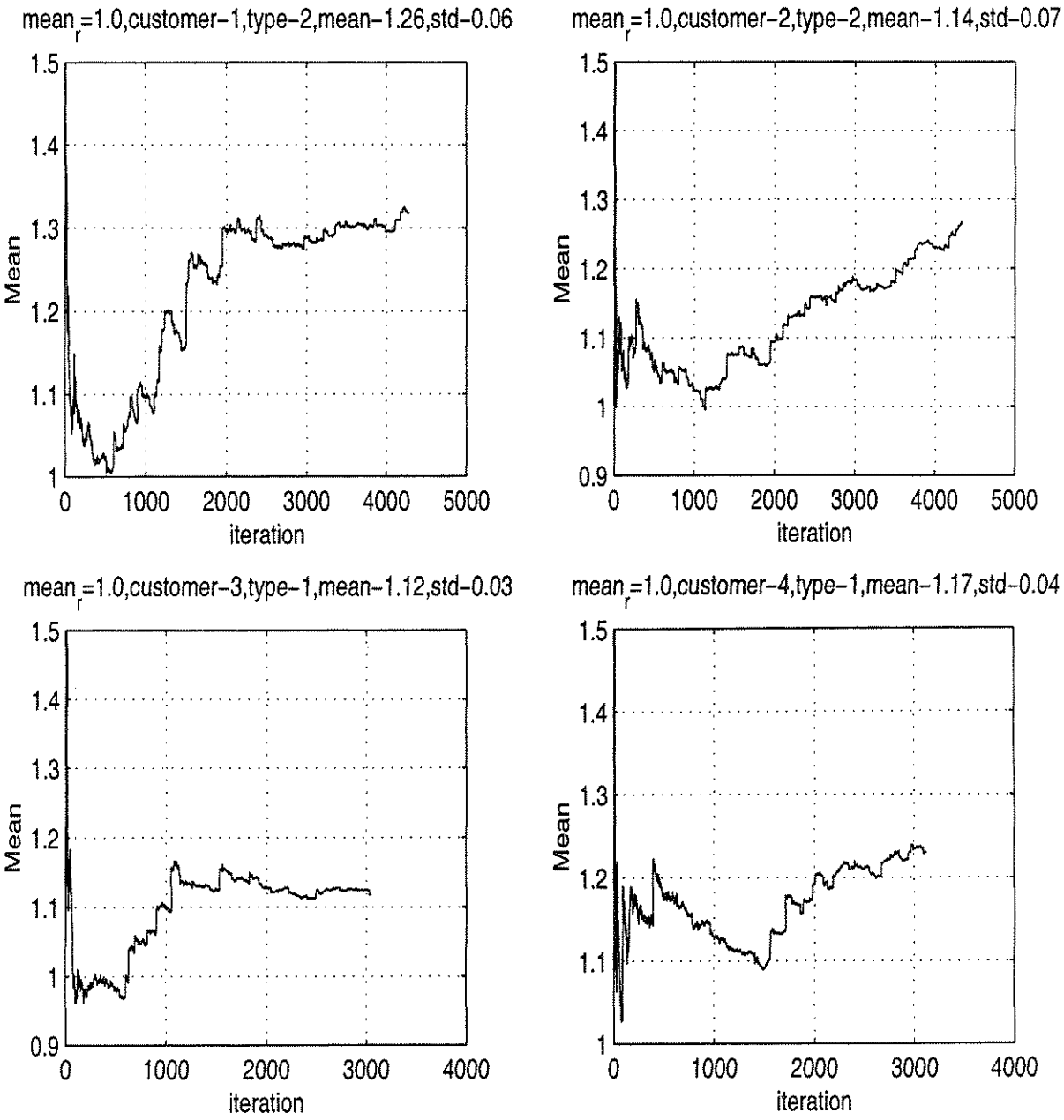

Figure 14: *Kaplan-Meier estimation of the waiting time, customers 1-4*

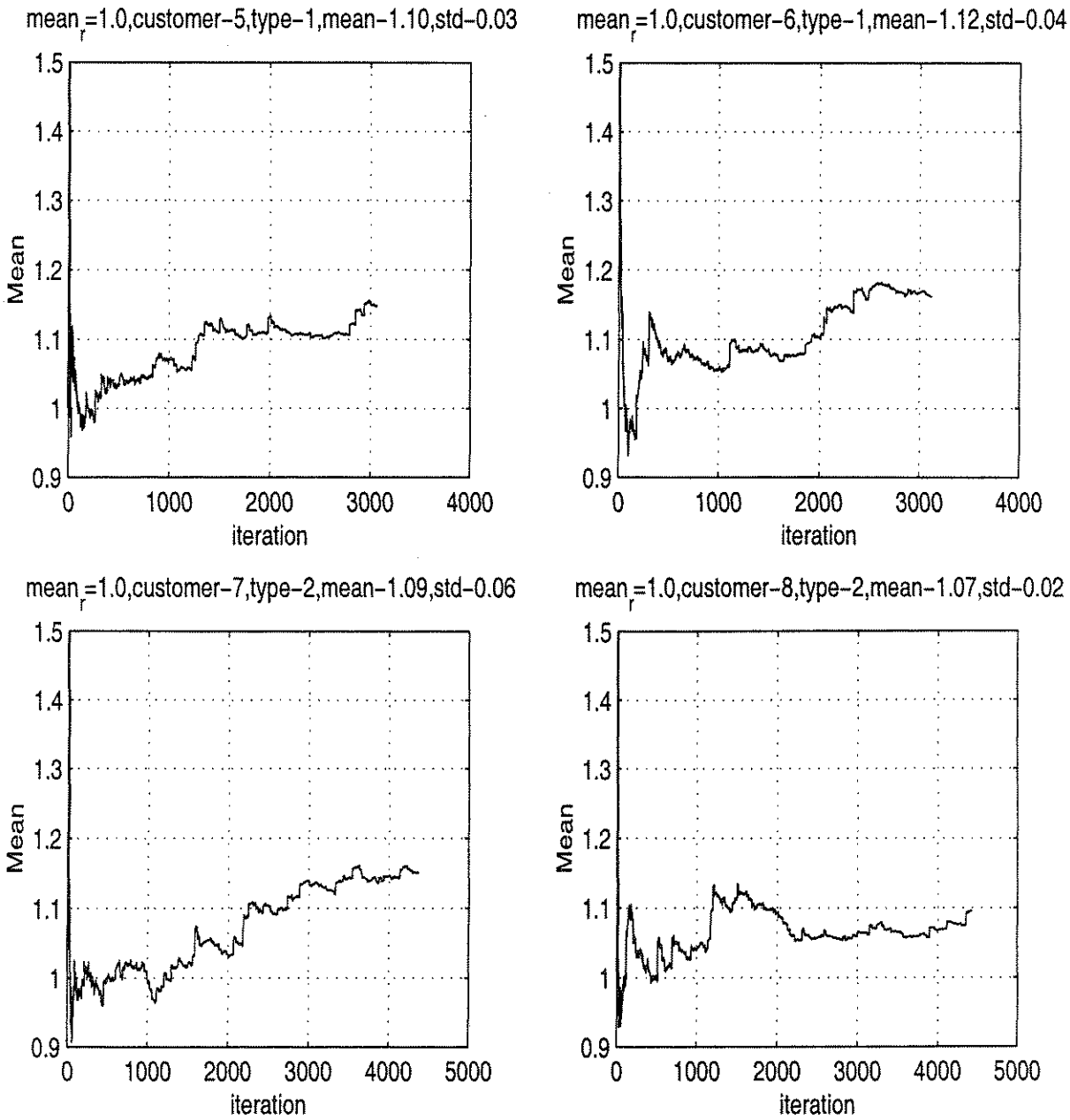

Figure 15: *Kaplan-Meier estimation of the waiting time, customers 5-8*

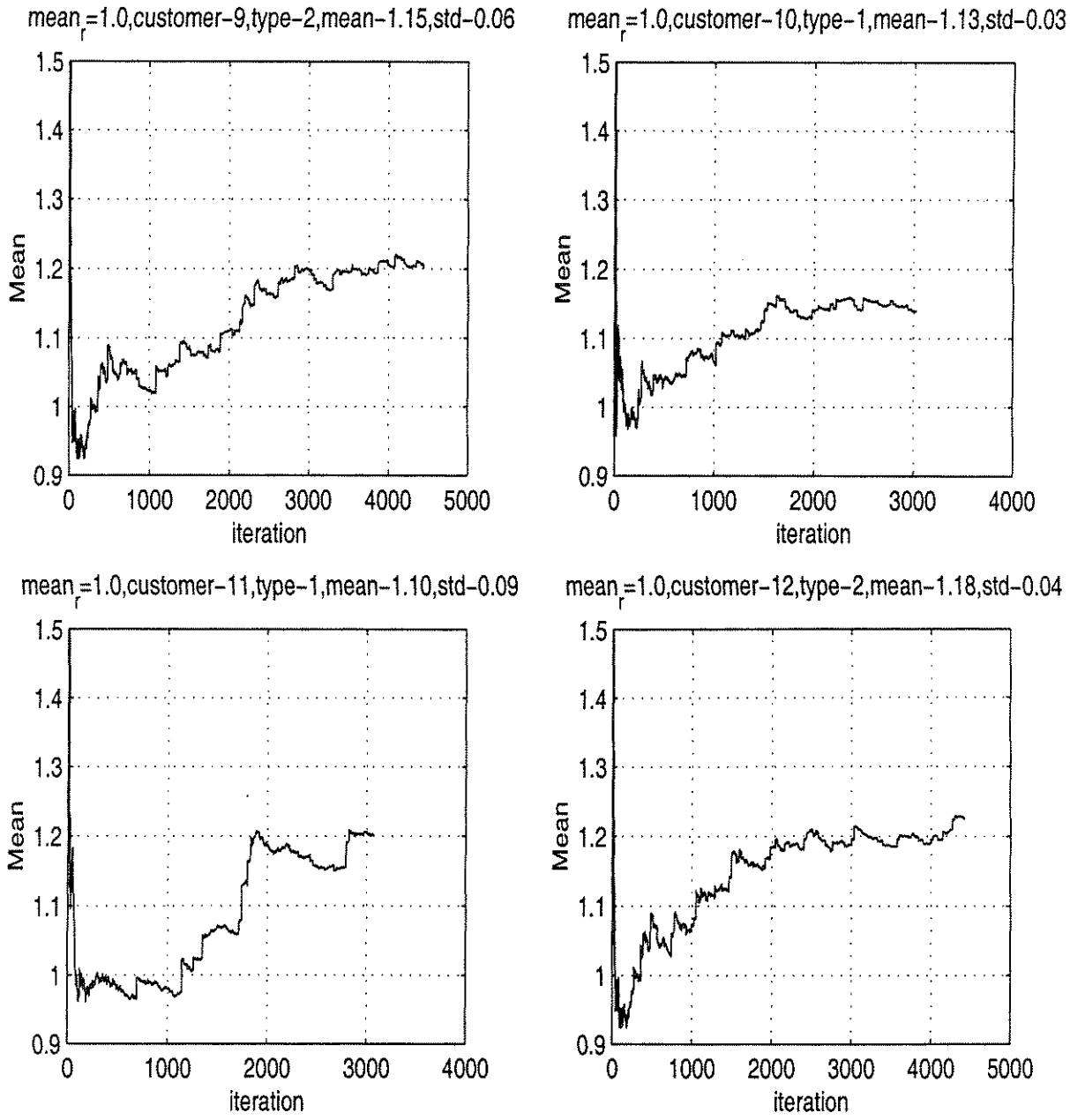

Figure 16: *Kaplan-Meier estimation of the waiting time, customers 9-12*

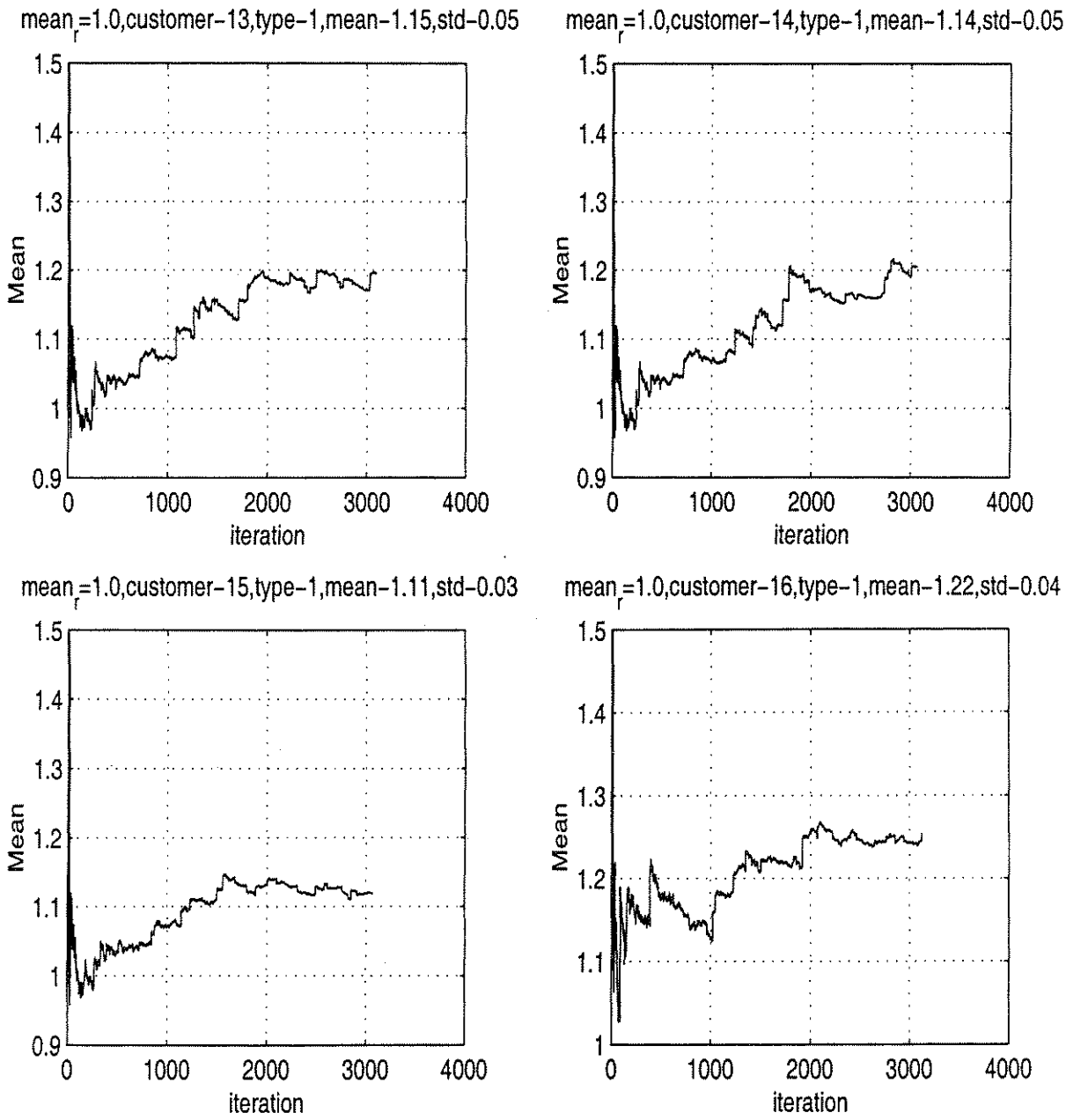

Figure 17: *Kaplan-Meier estimation of the waiting time, customers 13-16*

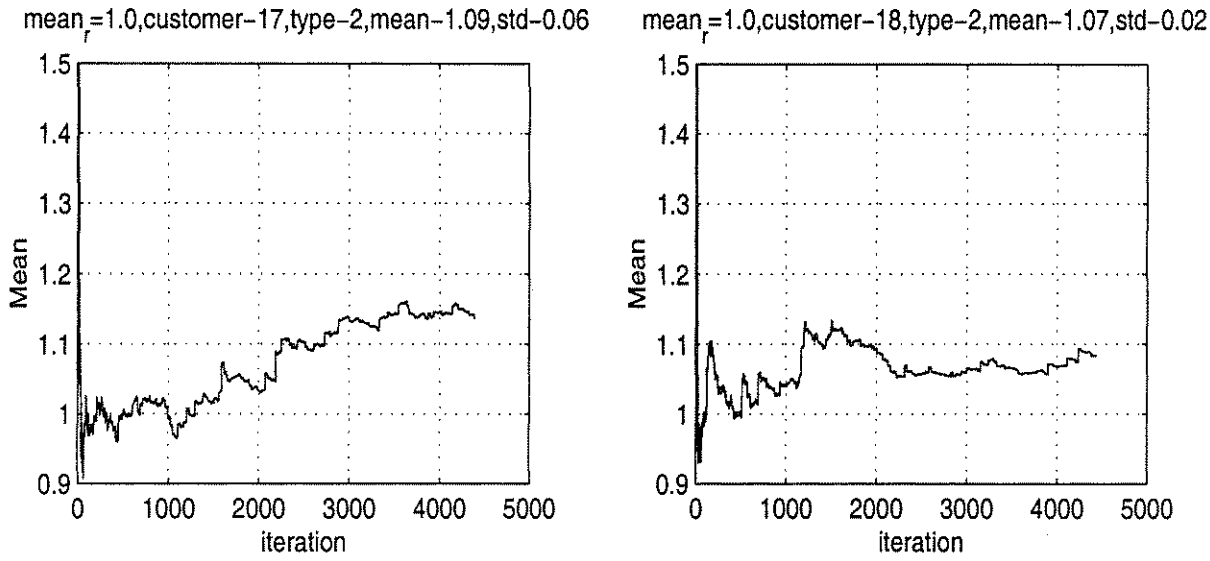

Figure 18: *Kaplan-Meier estimation of the waiting time, customers 17-18*

Example 4 The parameters are: $\lambda = 1$, $\mu = 1$, $T_0 = 1.5$, $Z = 1$, exploration period = 30, $T_1(x) = 0.95 * x$, $T_2(x) = 0.65 * x$, Censored MLE estimator.

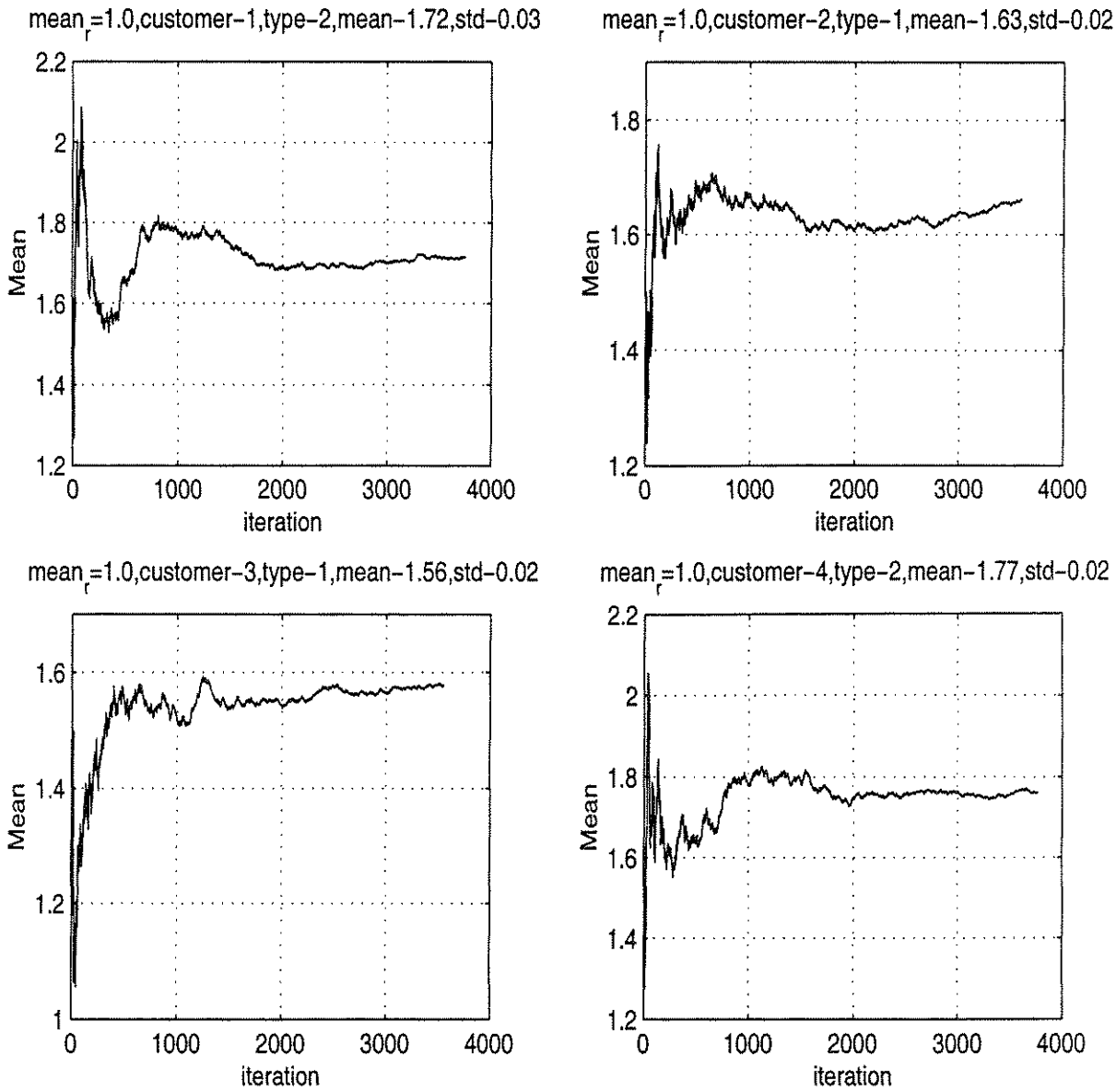

Figure 19: *Censored MLE estimation of the waiting time 1-4*

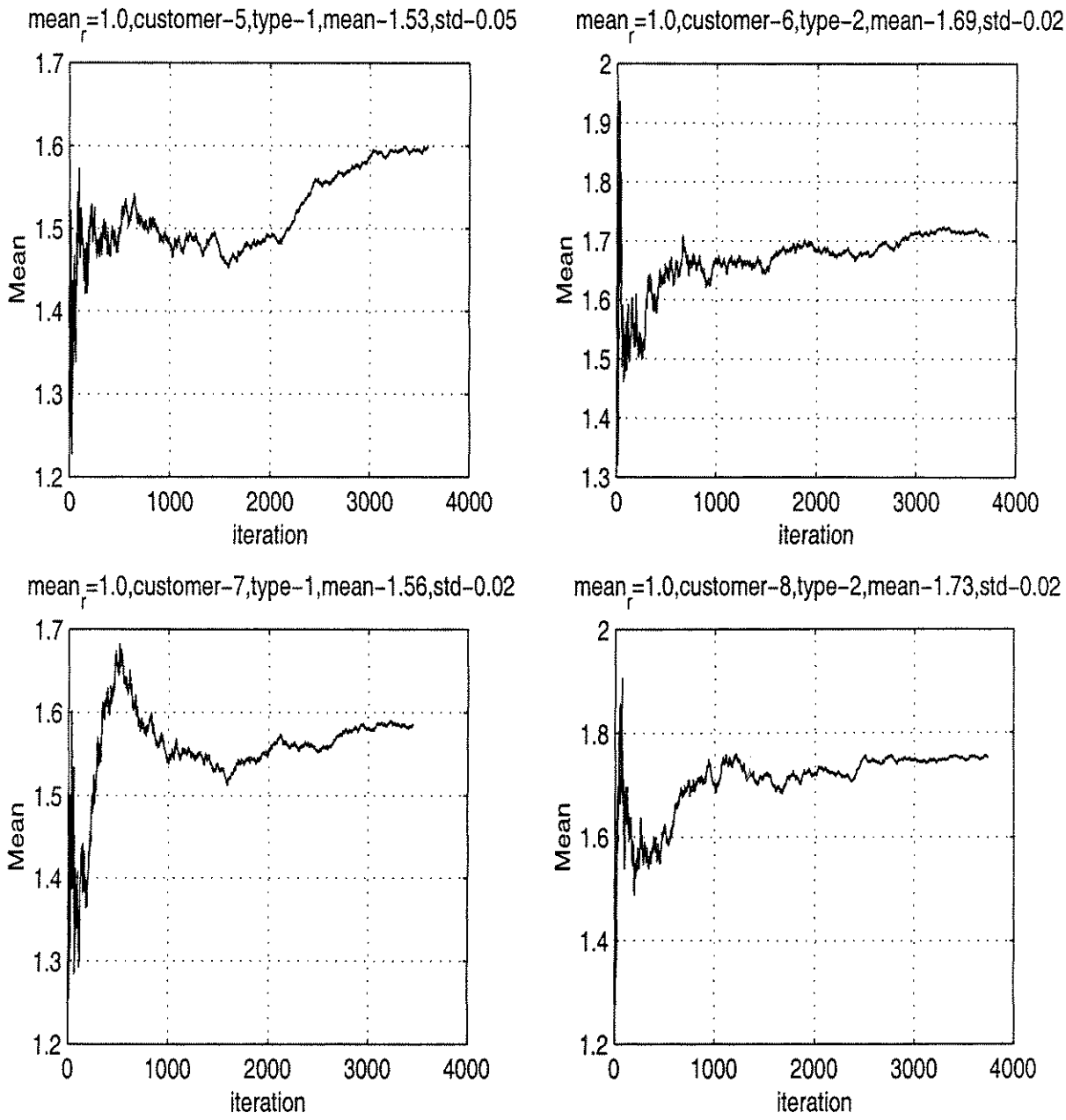

Figure 20: *Censored MLE estimation of the waiting time 5-8*

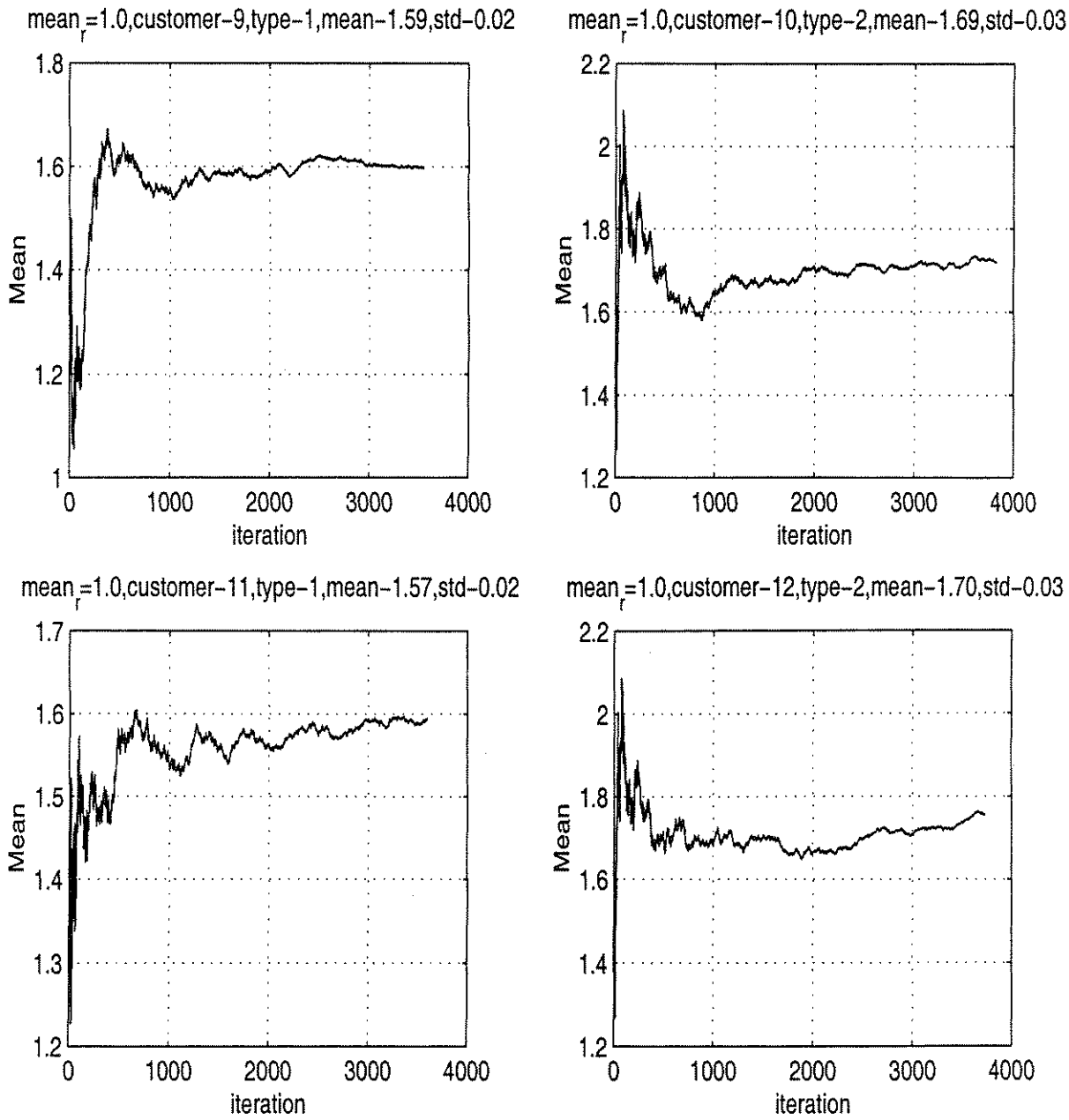

Figure 21: *Censored MLE estimation of the waiting time 9-12*

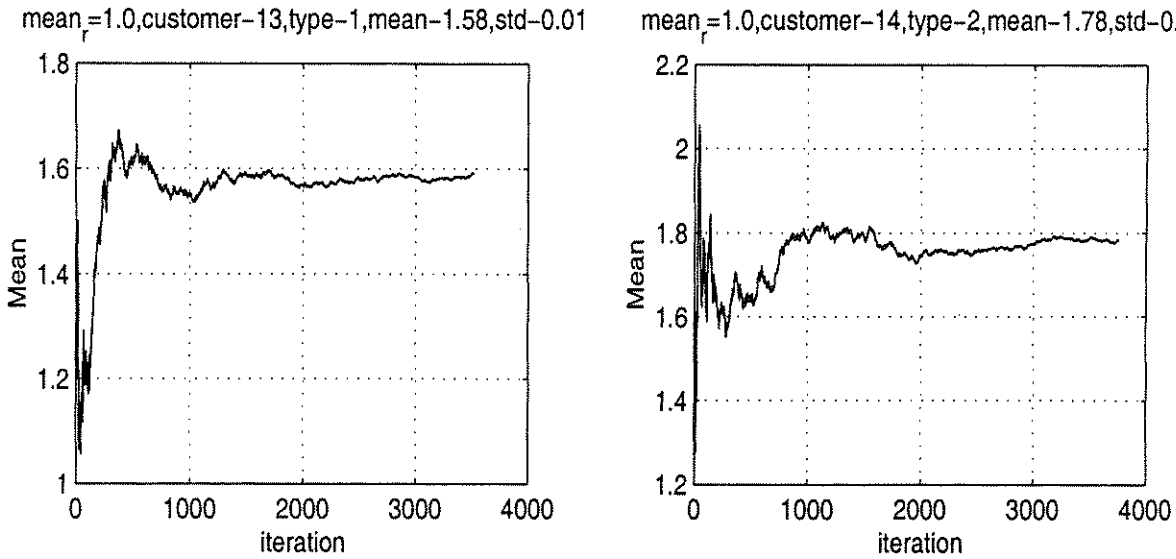

Figure 22: *Censored MLE estimation of the waiting time 13-14*

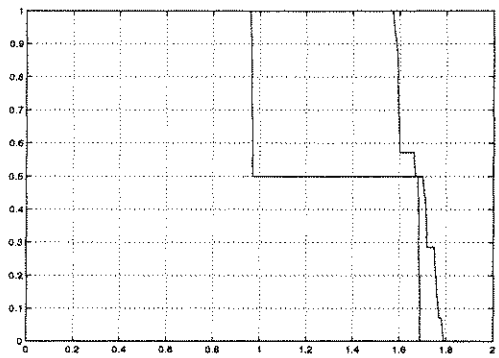

Figure 23: *The theoretic and real patience function*

Figure 23 shows the theoretic patience function of two types of customers (abandonment after T_1 and T_2) in comparison to the real patience calculated from the simulation results (see (5.9)).

The mean waiting time of all 14 customers is 1.6494 with standard deviation of 0.0848.
The mean waiting time of type 1 customers is 1.5741 with standard deviation of 0.0329.
The mean waiting time of type 2 customers is 1.7246 with standard deviation of 0.0359.
Now there are two abandonment/service ratios, each for every type.
The theoretic conditional calculated waiting time for two types of customers uses:

$$\bar{G}(x,t) = \begin{cases} 1 \text{ for } t \leq T_1 \\ 0.5 \text{ for } T_1 \leq t \leq T_2 \\ 0 \text{ for } t \geq T_2 \end{cases} \tag{8.20}$$

Type 1 ratio is:

$$x = \frac{\sum_{i=1}^{m} x_i}{m} + \frac{\sum_{j=1}^{n-m} T_j}{m} = \frac{0.95x}{2} + \frac{0.95(n-m)x}{m}$$

and the second calculation of the ratio is:

$$\frac{P(t > T_2)}{P(t < T_2)} = \frac{\exp(\frac{1}{2}(T_1 + T_2)) \int_{T_2}^{\infty} \exp(-s)ds}{\int_{0}^{T_1} ds + \exp(\frac{1}{2}(T_1) \int_{T_1}^{T_2} \exp(-\frac{1}{2}s)ds} = \frac{\exp(-0.15x)}{1 + 0.65x - \exp(-0.15x)}$$

Comparing both expressions, we get: $x = 1.774$.
The same way, we can calculate the estimated mean waiting time of type 2:

$$x = \frac{\sum_{i=1}^{m} x_i}{m} + \frac{\sum_{j=1}^{n-m} T_j}{m} = \frac{0.65x}{2} + \frac{0.65(n-m)x}{m}$$

and the second calculation of the ratio is:

$$\frac{P(t > T_1)}{P(t < T_1)} = \frac{\exp(\frac{1}{2}(T_1)) \int_{T_1}^{T_2} \exp(-\frac{1}{2}s)ds + \exp(\frac{1}{2}(T_1 + T_2)) \int_{T_2}^{\infty} \exp(-s)ds}{\int_{0}^{T_1} ds} = \frac{1}{0.65x}$$

The last equation results in $x = 1.482$.
Both theoretic values are close to the simulation values.

9 Conclusion

In this work we studied the behavior of an invisible M/M/m queue with abandonment. Our main assumption was that the dominant factor in how customers perceive system performance is their estimated mean waiting time. This affects customrs' decision of whether to stay or abandon the queue, determining, in turn, the system's load. This inter-relation between the system's performance and the customers' behaviour was addressed.

Two sets of customer's patience functions were explored. The first was the set of all functions decreasing in the estimated mean waiting time, x, and the other set consisted of all functions increasing in x, in a linearly bounded way. We have established that if the patience function decreases in x, there exists an equilibrium, where x equals the actual mean waiting time, and that this equilibrium is unique. For patience functions which are increasing in x, in a linearly bounded way, we have shown that the equilibruim is unique given that it exists.

We have suggested a rational decision model which compares a cost function with the hazard-rate function to establish the abandonment rule. Psychological research, which we have reviewed, motivates the assumption that the cost function is convex increasing in t. Additionally, since no information is available to the customers regarding the waiting time distribution, the customers were assumed to believe in an exponentially distributed waiting time. We obtained existence and uniqueness of equilibrium for the rational decision model by showing that the aggregate behavior of customers according to this decision rule yields patience functions within the explored sets.

We finally simulated the system for the case where customers patience is linearly dependent in x. The customers used their past experience to estimate the mean waiting time in the system using a censored sampling estimator. We have shown that if customers use the Kaplan-Meier estimator, a convergence to the partially consistent equilibruim prevails. On the other hand, if customers use the Censored MLE estimator, their false assumption of exponentialy distributed waiting time yields a bias. The simulation bias was found to

coincide with the theoretically anticipated bias.

In order to develop accurate models, more empirical data must be obtained and processed. Throughout this work we have assumed existence of a single parameter shaping customers patience. The main reason for settling with such an assumption is lack of data and experiments directed to reveal more parameters and their influence on customers. Gaining more knowledge is necessary for reducing the speculative part of this work.

Another area in which more work needs to be done is the visible or invisible + information queues analysis. By concentrating on invisible queues we eliminated the psychological aspect of interaction according to the information supplied by the system. Customers do not know in what place they are in the queue, when a new customer is admitted into service, whether the server is working, etc. Such information can play a major part in deciding if and when to abandon the queue.

Further research is also required in chracterizing different groups of customers. Customers can be distinguished according to their motivation to achieve the service [27], their willingness to pay in order to shorten the waiting time [7], their social status [4], and so on. Although there are some references that study different behavior versus group relation, a more extensive research is required. We believe that the principles presented in this work provide a foundation on which more accurate and strong models may be built.

10 Appendix – Simulation code

10.1 LRNSYS.M

```
% In order to change the cost function, changes must take place
% inside invgmat.m, Tz_APrmE.m and Tz_APrmF.m

global MLE APrmE APrmF MLENC FileNum initTz T0 ThrowExtrm DumpThrsh DumpFactor
ExplorePrd alfa

OutMat=[];%%%% Parameters to be simulated %%%%
%MLE3_8_1_5;
%%%% first service and arrival times %%%%
rand('state',sum(100*clock));
u= rand(1,1);
dtserv=(-1/mu)*log(u); %service time
u= rand(1,1);
dtarv=(-1/lamda)*log(u); %arrival time
tarv=tarv+dtarv;
tserv=tarv+dtserv;

%profile on
%%%% Queue loop %%%%
for step=1:N
  if (mod(step,500)==0)
    save count step
  end

  f=1;
  if (size(ClientsMat)~=0)
    for o=1:length(ClientsMat(:,3))
      if (ClientsMat(o,3)==length(DataMat{o,1}))
        f1(o)=0;
      end
    end
  end


  %updating queue list
  if (tserv>=tarv) %next client arrived
    QueueLen=size(QueueMat,1);
    if (QueueLen>1) %queue not empty (one client is served and the rest are waiting)
      [QueueMat,DataMat,ClientsMat]= ...
        cln_que(QueueMat,ClientsMat,tarv,DataMat,EstimFlag);
    end

    %%%% client data %%%%
    % finding the client's type
    u= rand(1,1); zcurr=length(find(CliIntrv<u)); %choosing client type
    if (zcurr==0) zcurr=CliNum; end

    % finding the client's number and updatingClientsMat if needed
    [clicurr,ClientsMat,DataMat]=client_z(ClientsMat,DataMat,zcurr,EstimFlag);
```

```
    %client enters the queue
    QueueLen=size(QueueMat,1);
    if (QueueLen==0)
      QueueMat(QueueLen+1,:)=[clicurr, tarv, tarv];
    else     % time of entering service will be updated later
      QueueMat(QueueLen+1,:)=[clicurr, tarv, 0];
    end

    %next arrival time
    u= rand(1,1);
    dtarv=(-1/lamda)*log(u); %arrival time
    tarv=tarv+dtarv;

  else  %next client served (FCFS)
    u= rand(1,1);
    dtserv=(-1/mu)*log(u); %service time

    QueueLen=size(QueueMat,1);
    if (QueueLen>=1) %queue not empty
      if (QueueLen>1) %queue not empty (one client is served and the rest are waiting)
        [QueueMat,DataMat,ClientsMat]= ...
          cln_que(QueueMat,ClientsMat,tserv,DataMat,EstimFlag);
      end

      %update data for serviced client
      QueueLen=size(QueueMat,1);
      [ClientsMat,DataMat]= ...
        updt_cli(QueueMat(1,:),ClientsMat,DataMat,EstimFlag);

       if (QueueLen>1)
        QueueMat=QueueMat(2:QueueLen,:);
        QueueMat(1,3)=tserv;
      else QueueMat=[];
      end
      tserv=tserv+dtserv;
    else
      tserv=tarv+dtserv;
    end
  end
  Len=[Len, size(QueueMat,1)];
end % Queue loop

%profile report
pltgrph=0;
if pltgrph==1

len=length(Len);
estmean=mean(Len(len-1000:len));
eststd=std(Len(len-1000:len));
S=sprintf('Size of Queue \nreal mean=%2.1f, est. std-%2.3f, est. mean-%2.3f',1/mu,eststd,estmean);
plot(Len,'b');title(S);xlabel('iteration');ylabel('Size');
grid;
zoom on;

pltgrph=0;
estmeanvec=[];
```

```
for k=1:size(ClientsMat,1)
  res=mod(k,4);
  if (res==0) res=4; end
  len=length(DataMat{k,3});
  r=round(0.8*len);
  estmean=mean(DataMat{k,3}(len-r:len));
  estmeanvec=[estmeanvec, estmean];
  eststd=std(DataMat{k,3}(len-r:len));
  OutMat(k,:)=[k, ClientsMat(k,1), estmean, eststd];
  if pltgrph==1
    if (res==1) figure, end

S=sprintf('mean_r=%2.1f,client-%d,type-%d,mean-%2.1f,std-%2.2f',1/mu,k,ClientsMat(k,1),estmean,estst
d);
    subplot(2,2,res);
    if ((EstimFlag==MLE)|(EstimFlag==MLENC))
     plot(1:len,DataMat{k,3});title(S);xlabel('iteration');ylabel('Mean');
    else
     Ptlen=length(DataMat{k,4});
     % calculating St
     St=DataMat{k,4}(1:Ptlen);
     for m=2:Ptlen
       St=St.*[ones(1,(m-1)), (DataMat{k,4}(1:(Ptlen-m+1)))];
     end
     plot([0, DataMat{k,1}],[1, St]);title(S);xlabel('time');ylabel('St');
    end
    grid;
    zoom on;
  end
end
end

if EstimFlag==APrmF
  figure;
  plot(ClientsMat(:,6),'o');grid on; zoom on;
end
```

10.2 CLN_QUE.M

```
% Cleans the queue from clients that by time - t have already left.

function [QueueMat,DataMat,ClientsMat]=cln_que(QueueMat,ClientsMat,t,DataMat,EstimFlag)

global MLE APrmE APrmF MLENC

ServCli=QueueMat(1,:);  % save served client
newTi=[];
[QueueLen,J]=size(QueueMat);
QueueMat=QueueMat(2:QueueLen,:);  %clean the waiting clients only

vec=QueueMat(:,2)+ClientsMat(QueueMat(:,1),6);
[I,J]=find(vec<t); %abandon
l=length(I);
if (l~=0)
 %update abandonment data
 for k=1:l
   newTi(k,1)=ClientsMat(QueueMat(I(k),1),6);
   if ((EstimFlag~=MLE)&(EstimFlag~=MLENC))
     len=length(DataMat{QueueMat(I(k),1),2});
     if (len==0)
       DataMat{QueueMat(I(k),1),2}=newTi(k,1);
     else
       indx=length(find(DataMat{QueueMat(I(k),1),2}<newTi(k,1)));
       %new Ti
       DataMat{QueueMat(I(k),1),2}= ...
       [DataMat{QueueMat(I(k),1),2}(1:indx),newTi(k,1),DataMat{QueueMat(I(k),1),2}(indx+1:len)];
     end
   end
 end
 % z   in=1  Nx Sumx SumT Tz Nt
 ClientsMat(QueueMat(I,1),2)=0; %in=0
 ClientsMat(QueueMat(I,1),5)=ClientsMat(QueueMat(I,1),5)+ClientsMat(QueueMat(I,1),6);%SumT
 ClientsMat(QueueMat(I,1),7)=ClientsMat(QueueMat(I,1),7)+1;%Nt
 switch EstimFlag,   %update Tz according to the right estimator
  case MLE,
      [ClientsMat(QueueMat(I,1),6),DataMat(QueueMat(I,1),:)]= ...
        Tz_MLE(ClientsMat(QueueMat(I,1),:),DataMat(QueueMat(I,1),:),EstimFlag);
  case MLENC,
      [ClientsMat(QueueMat(I,1),6),DataMat(QueueMat(I,1),:)]= ...
        Tz_MLE(ClientsMat(QueueMat(I,1),:),DataMat(QueueMat(I,1),:),EstimFlag);
  case APrmE,
      [ClientsMat(QueueMat(I,1),6),DataMat(QueueMat(I,1),:)]= ...
        Tz_APrmE(ClientsMat(QueueMat(I,1),:),DataMat(QueueMat(I,1),:),newTi,0);
 end
end

[I,J]=find(vec>=t); %still waiting
if (length(I)==0) QueueMat=[];
else QueueMat=QueueMat(I,:);
end
```

```
QueueLen=size(QueueMat,1);
% place the served client back in the queue
if (QueueLen==0)
  QueueMat=ServCli;
else
  QueueMat=[ServCli; QueueMat];
end
```

10.3 CLIENT_Z.M

```
% Chooses the client number, given the client type and the queue status.
% There are three possibilities:
% (1) new type in the queue->initialize new client type.
% (2) new client from existing type->new client with old client information.
% (3) free old client->reentering old client

function [clicurr,ClientsMat,DataMat]=client_z(ClientsMat,DataMat,zcurr,EstimFlag)

global MLE APrmE APrmF MLENC initTz

CliMatLen=size(ClientsMat,1);
if (CliMatLen==0) I=[];
else
  I=find(ClientsMat(:,1)==zcurr);
  if (length(I)~=0)  %there are/were zcurr clients in the queue
    l=find(ClientsMat(I,2)==0);
    if (length(l)~=0) J=I(l); else J=[]; end %check for possible clients
    tmp=length(J);
  end
end

if (length(I)==0) %no clients from this type ever in the queue (1)
  clicurr=CliMatLen+1;
  if ((EstimFlag~=MLE)&(EstimFlag~=MLENC))
     ClientsMat(clicurr,:)=[zcurr, 1,  10,  10*initTz,  0, initTz, 0];
     %  Pt - Calculation for 10 initial values
     DataMat{clicurr,4}=[0.9, 0.9*(1-1/9), 0.9*(1-1/9)*0.875, 0.9*(1-1/9)*0.875*(1-1/7),...
                  0.9*(1-1/9)*0.875*(1-1/7)*(1-1/6), 0.9*(1-1/9)*0.875*(1-1/7)*(1-1/6)*0.8,...
                  0.9*(1-1/9)*0.875*(1-1/7)*(1-1/6)*0.8*0.75, ...
                  0.9*(1-1/9)*0.875*(1-1/7)*(1-1/6)*0.8*0.75*(1-1/3), ...
                  0.9*(1-1/9)*0.875*(1-1/7)*(1-1/6)*0.8*0.75*(1-1/3)*0.5, 0];
     %          Xi
     DataMat{clicurr,1}=[0.95*initTz, 0.96*initTz, 0.97*initTz, 0.98*initTz, 0.99*initTz, ...
                  initTz, 1.01*initTz, 1.02*initTz, 1.03*initTz, 1.04*initTz];
     DataMat{clicurr,3}=initTz*ones(1,10);  % E
     DataMat{clicurr,2}=[];      % Ti
  else
                   % z   in=1  Nx   Sumx  SumT   Tz    Nt
     ClientsMat(clicurr,:)=[zcurr, 1,  10,  10*initTz,  0, initTz, 0];
     %          Xi
     DataMat{clicurr,1}=[0.95*initTz, 0.96*initTz, 0.97*initTz, 0.98*initTz, 0.99*initTz, ...
                  initTz, 1.01*initTz, 1.02*initTz, 1.03*initTz, 1.04*initTz];
     DataMat{clicurr,2}=[]; % Ti
     DataMat{clicurr,3}=initTz*ones(1,10); %  E

  end

  elseif (tmp==0) %same type clients in the queue    (2)
  sametypecli=length(I);
  tmpvec=(1:sametypecli)/sametypecli;
  u= rand(1,1);
  k=length(find(tmpvec<u));
  if (k==0) k=sametypecli; end
```

```
  clicurr=I(k); %number of new client in the queue
  clinew=CliMatLen+1;
                  % z     in=1 Nx Sumx SumT Tz
  ClientsMat(clinew,:)=[ zcurr,   1, ClientsMat(clicurr,3:7)];
  DataMat{clinew,1}=DataMat{clicurr,1};
  DataMat{clinew,2}=DataMat{clicurr,2};
  DataMat{clinew,3}=DataMat{clicurr,3};
  if ((EstimFlag~=MLE)&(EstimFlag~=MLENC)) DataMat{clinew,4}=DataMat{clicurr,4}; end
  clicurr=clinew;

else %free client from zcurr type   (3)
  tmpvec=(1:tmp)/tmp;
  u= rand(1,1);
  k=length(find(tmpvec<u));
  if (k==0) k=tmp; end
  clicurr=I(l(k)); %number of new client in the queue
  ClientsMat(clicurr,2)=1;
end
```

10.4 UPDT_CLI.M

```
% Updates the current client data

function [ClientsMat,DataMat]= ...
        updt_cli(QueueVec,ClientsMat,DataMat,EstimFlag)

global MLE APrmE APrmF MLENC

clinum=QueueVec(1,1);
tarvque=QueueVec(1,2);
tentrs=QueueVec(1,3);
newXi=(tentrs-tarvque);
if newXi<0
  disp('newXi<0');
  pause;
elseif newXi>0
 if ((EstimFlag~=MLE)&(EstimFlag~=MLENC))
   len=length(DataMat{clinum,1});
   if (len==0)
     DataMat{clinum,1}=newXi;
   else
     indx=length(find(DataMat{clinum,1}<newXi));
     %new Xi
     DataMat{clinum,1}=[DataMat{clinum,1}(1:indx),newXi,DataMat{clinum,1}(indx+1:len)];
   end
 end

        % z   in=1  Nx Sumx SumT Tz
 ClientsMat(clinum,2)=0; %in=0
 ClientsMat(clinum,3)=ClientsMat(clinum,3)+1;  %Nx
 ClientsMat(clinum,4)=ClientsMat(clinum,4)+newXi;  %SumX
 %calculating the abandon time
 switch EstimFlag,    %update Tz according to the right estimator
  case MLE,
   [ClientsMat(clinum,6),DataMat(clinum,:)]= ...
     Tz_MLE(ClientsMat(clinum,:),DataMat(clinum,:),EstimFlag);
  case MLENC,
   [ClientsMat(clinum,6),DataMat(clinum,:)]= ...
     Tz_MLE(ClientsMat(clinum,:),DataMat(clinum,:),EstimFlag);
  case APrmE,
   [ClientsMat(clinum,6),DataMat(clinum,:)]= ...
     Tz_APrmE(ClientsMat(clinum,:),DataMat(clinum,:),newXi,1);
 end
else
 ClientsMat(clinum,2)=0; %in=0
end
```

10.5 Tz_MLE.M

```
% Updates the saved Tz data for the given clients
% exp. distribution assumption
% Maximum Likelihood Estim. using censored sampling

function [Tz,DataMat]=Tz_MLE(ClientsMat, DataMat, EstimFlag)

global MLE MLENC T0 ExplorePrd alfa dalfa rndU

if (EstimFlag==MLE)
  %MLE estimator Censored samples
  E=(ClientsMat(:,4)+ClientsMat(:,5))./(ClientsMat(:,3)+eps*ones(size(ClientsMat(:,4))));
else
  %MLE estimator service samples
  E=(ClientsMat(:,4))./(ClientsMat(:,3)+eps*ones(size(ClientsMat(:,4))));
end

l=size(ClientsMat,1);
for k=1:l
   DataMat{k,3}=[DataMat{k,3}, E(k)]; %new E
end

%calculating the abandon time <----- update when changing T(E)
Prd=(ClientsMat(:,3)+ClientsMat(:,7))/ExplorePrd;
[I,J]=find(round(Prd)==Prd);
if (length(I)~=0)
   Tz(I,1)=1/eps*ones(size(I));
   [I,J]=find(round(Prd)~=Prd);
   %Tz(I,1)=E(I,1)./ClientsMat(I,1)+1./E(I,1); %abandonment rule
   if (length(I)~=0)
      if (rndU==1)
        U=0.01*randn(size(I));
        U(find(U<-0.2))=zeros(size(find(U<-0.2)));
      else
        U=zeros(size(I));
      end
      Tz(I,1)=alfa*E(I,1)-dalfa*(ClientsMat(I,1)-ones(size(I))).*E(I,1)+U; %abandonment rule
      %[K,L]=find(Tz(I)>T0);
      %if (length(L)~=0) Tz(I(L),1)=T0*ones(size(I(L))); end  %Tz=T0
   end
else
   %Tz=E./ClientsMat(:,1)+1./E; %abandonment rule
   U=0.01*randn(size(E));
   U(find(U<-0.2))=zeros(size(find(U<-0.2)));
   Tz=alfa*E-dalfa*(ClientsMat(:,1)-ones(size(E))).*E+U; %abandonment rule
   %[K,L]=find(Tz>T0);
   %if (length(K)~=0) Tz(K)=T0*ones(size(K)); end  %Tz=T0
end
```

10.6 Tz_APRME.M

```
% Updates the saved Tz data for the given clients
% aparametric estimation of the mean waiting time in the queue
% newF=1 if newV is Xi value

function [Tz,DataMat]=Tz_APrmE(ClientsMat, DataMat, newV, newF)

global T0 ExplorePrd alfa dalfa rndU

E=[];
% update S(t) data
[DataMat, ClientsMat]=updt_Pt(ClientsMat,DataMat,newV,newF);

l=size(ClientsMat,1);
for k=1:l
  Ptlen=length(DataMat{k,4});
  if Ptlen>0
    % calculating St
    St=DataMat{k,4}(1:Ptlen);
    for m=2:Ptlen
      St=St.*[ones(1,(m-1)), (DataMat{k,4}(1:(Ptlen-m+1)))];
    end
    % calculating the mean waiting time
    E(k,1)=trapz([0,DataMat{k,1}],[1, St]);
    DataMat{k,3}=[DataMat{k,3}, E(k,1)]; %new E
  else
    E(k,1)=eps;
  end
end

%calculating the abandon time <----- update when changing T(E)
Prd=(ClientsMat(:,3)+ClientsMat(:,7))/ExplorePrd;
[I,J]=find(round(Prd)==Prd);
if (length(I)~=0)
  Tz(I,1)=1/eps*ones(size(I));
  [I,J]=find(round(Prd)~=Prd);
  %Tz(I,1)=E(I,1)./ClientsMat(I,1)+1./E(I,1); %abandonment rule
  if (length(I)~=0)
    if (rndU==1)
      U=0.01*randn(size(I));
      U(find(U<-0.2))=zeros(size(find(U<-0.2)));
    else
      U=zeros(size(I));
    end
    Tz(I,1)=alfa*E(I,1)-dalfa*(ClientsMat(I,1)-ones(size(I))).*E(I,1)+U; %abandonment rule
    %[K,L]=find(Tz(I)>T0);
    %if (length(L)~=0) Tz(I(L),1)=T0*ones(size(I(L))); end  %Tz=T0
  end
else
  %Tz=E./ClientsMat(:,1)+1./E; %abandonment rule
  U=0.01*randn(size(E));
  U(find(U<-0.2))=zeros(size(find(U<-0.2)));
  Tz=alfa*E-dalfa*(ClientsMat(:,1)-ones(size(E))).*E+U; %abandonment rule
  %[K,L]=find(Tz>T0);
  %if (length(K)~=0) Tz(K)=T0*ones(size(K)); end  %Tz=T0
end
```

10.7 UPDT_PT.M

```
% updates the distribution estimation (the mult. elements-Pt used to construct
% the estimation St)

function [DataMat, ClientsMat]=updt_Pt(ClientsMat,DataMat,newV,newF)

global ThrowExtrm DumpThrsh DumpFactor

l=size(ClientsMat,1);
for k=1:l
  calwhole=0;
  %thorwing extreme values of Xi and calculating the whole Pt
  if ((ThrowExtrm==1)&(DumpThrsh<=ClientsMat(k,3))) %filter data
    meanXi=mean(DataMat{k,1});
    thrwindx=ClientsMat(k,3)-length(find(DataMat{k,1}>DumpFactor*meanXi))
    if (thrwindx<ClientsMat(k,3))  %data is filtered
      ClientsMat(k,4)=ClientsMat(k,4)-sum(DataMat{k,1}((thrwindx+1):ClientsMat(k,3)));
      DataMat{k,1}=DataMat{k,1}(1:thrwindx);
      ClientsMat(k,3)=thrwindx;
      calwhole=1;
    end
  end
      %Nx
  indx=ClientsMat(k,3)-length(find(DataMat{k,1}>newV(k,1)));
  lenPt=length(DataMat{k,4});
  %data is filtered->calculate everything
  if (calwhole==1)
    indx=ClientsMat(k,3);
    DataMat{k,4}=[];
  %new Ti is smaller than any Xi->do nothing
  elseif ((indx==0)&(ClientsMat(k,3)>0)&(newF==0))
  %new Xi is smaller than any Xi->first element calculated in for loop and
  %inserted on the left of the Pt line
  elseif ((indx==1)&(newF==1))
  %new Ti or Xi is bigger than any Xi->calculate everything
  elseif (indx==ClientsMat(k,3))
    DataMat{k,4}=[];
  %new Xi is in the middle
  elseif ((indx>1)&(newF==1))
    indx=indx+1;
    DataMat{k,4}=DataMat{k,4}(indx:lenPt);
  %new Ti is in the middle
  else
    DataMat{k,4}=DataMat{k,4}(indx+1:lenPt);
  end

  %if newV is the first in the vector indx==0 if newV=Ti and Pt stays the same
  for j=indx:-1:1
    factor=length(find(DataMat{k,1}==(DataMat{k,1}(j))));
    n=length(find(DataMat{k,1}>=DataMat{k,1}(j)));
    m=length(find(DataMat{k,2}>=DataMat{k,1}(j)));
    DataMat{k,4}=[(1-(factor/(n+m))),DataMat{k,4}];
  end
end
```

References

[1] F. Baccelli and G. Hebuterne, "On queues with impatient customers," in F.J. Kylstra (ed.), *Performance '81*, North Holland, 1981, pp. 159–179.

[2] A.J. Brigandi, D.R. Dargon, M.J. Sheehan, and T. Spencer, "AT&T's call processing simulator (CPPS) operational design for in-bound call centers," Interfaces 24:1 (1994) 6–28.

[3] Z. Carmon and D. Kahneman, "The experienced utility of queuing: experience profiles and retrospective evaluations of simulated queues," preprint, 1998.

[4] A. Diekmann, M. Jungbauer-Gans, H. Krassnig, S. Lorenz, "Social status and aggression: A field study analyzed by survival analysis", Journal of Social Psychology, Dec 1996, pp. 761-768.

[5] Direct Marketing Association, "1999 Ecomomic Impact: U.S. Direct Marketing Today Executive Summary", http://www.the-dma.org/for_eric/DMAi/library/publications/libres-ecoimpact2.shtml

[6] L. Dube, B. H.Schmitt, F. Leclerc, "Consumers' Affective Response to Delays at Different Phases of a Service Delivery", Journal of Applied Social Psychology, 1991, 21, pp. 810-820.

[7] H. H. Friedman, L. W. Friedman, "Reducing the 'wait' in waiting-line systems: Waiting line segmentation", Magazine: Business Horizons, July/August 1997.

[8] T. Gail, L. Scott, "A field study investigating the effect of waiting time on customer satisfaction", Journal of Psychology, Nov 1997, pp. 655-660.

[9] O. Garnet and A. Mandelbaum, "Design of large call centers with impatient customers," preprint, 1998.

[10] R. Hassin and M. Haviv, "Equilibrium strategies for queues with impatient customers," Operations Research Letters 17 (1995) 41–45.

[11] R. Haviv and Y. Ritov, "Homogeneous customers renege at random times when waiting conditions deteriorate"

[12] J. Hornik, "Subjective Vs. Objective Time Measures: A Note on the Perception of Time in Consumer Behavior", Journal of Consumer Research, Vol. 11, June 1984.

[13] J. Hueter, W. Swart, "An Integrated Labor-Management System for Taco Bell", Institute for Operations Research, Jan-Feb 1998, pp. 75-91.

[14] M.K. Hui and D.K. Tse, "What to tell customers in waits of different lengths: an iterative model of service evaluation," Journal of Marketing 60 (1996) 81–90.

[15] J. G. Kalbfleisch, "Probability ans Statistical Inference", Springer-Verlag, 1979.

[16] N. M. Kiefer, "Economic duration data and hazard functions", Journal of economic literature, Vol. XXVI, June 1988, pp. 646-679.

[17] R. C. Larson, "Perspectives on queues: social justice and the psychology of queueing", Operation Research, Nov-Dec 1987, pp. 895-905.

[18] D. H. Maister, "The Psychology of Waiting Lines", Harvard Business School, 1984.

[19] A. Mandelbaum and N. Shimkin, "A model for rational abandonment from invisible queues", Queueing Systems: Theory and Applications (QUESTA), Vol 36, Nov 2000, pp. 141-173.

[20] E.E. Osuna, "The psychological cost of waiting", Journal of mathematical psychology, 29 (1985), pp.82-105

[21] C. Palm, "Methods of judging the annoyance caused by congestion", *Tele*, No. 2 1953, pp. 1–20.

[22] M. K. B. Parmar, D. Machin, "Survival Analysis: A Practical Approach", John Wiley & Sons, 1995.

[23] M. Shaked, J. G. Shanthikumar, "Stochastic Orders and their Applications", ACADEMIC PRESS, 1994.

[24] R. Suck, H. Holling, "Stress caused by waiting: a theoretical evaluation of a mathematical model", Journal of mathematical psychology 41 (1997), pp. 280-286.

[25] S. Taylor, "Waiting for Service: The Relationship Between Delays and Evaluations of Service", Journal of Marketing, April 1994, pp. 56-69.

[26] S. Taylor, J. D. Claxton, "Delays and the Dynamics of Service Evaluations", Journal of the academy of Marketing Science, Vol. 22, No. 3, 1994, pp. 254-264.

[27] M. Thierry, "Subjective importance of goal and reactions to waiting in line", Journal of Social Psychology, Dec 1994, pp. 819-827.

התנהגות מסתגלת של לקוחות חסרי סבלנות בתורים בלתי נראים

חיבור על מחקר
לשם מילוי חלקי של הדרישות לקבלת תואר
מגיסטר למדעים בהנדסת חשמל

אסתר זוהר

הוגש לסנט הטכניון - מכון טכנולוגי לישראל
שבט תשס״א חיפה ינואר 2001

המחקר נעשה בהנחיית פרופסור נחום שימקין ופרופסור אבישי מנדלבאום
בפקולטה להנדסת חשמל.

ברצוני להודות לפרופסור נחום שימקין ולפרופסור אבישי מנדלבאום על התמיכה
הרבה והסיוע שנתנו לי במהלך השתלמותי. לולא עזרתם הרבה לא היתה עבודה זו
מגיעה לכלל סיום.

אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

תוכן עניינים

תקציר **1**

רשימת סמלים **3**

הקדמה **4**

רקע ומוטיבציה 4
הנחות ותוצאות 5
מודל למידה – תיאור וסימולציה 6

סקר ספרות **8**

סקר ספרות – מערכות תורים 8
סקר ספרות – פסיכולוגיה 9

תיאור המודל **14**

מודל המערכת 14
פונקצית הסבלנות 16
עזיבה רציונלית 23

פונקצית התועלת 23

פונקצית המחיר 25
תכונות ההחלטה האופטימלית 28

שיווי משקל ויחידות **30**

ניתוח השרדות **33**

קטימה בשיערוך זמן ההמתנה 33
משערך קפלן – מאייר 34
משערך ניראות מירבית קטום 35

מודל למידה – תיאור וסימולציה **36**

מסקנות **64**

רשימת מקורות **66**

רשימת איורים

איור 1 : שיערוך פילוג זמן ההמתנה לקוחות 4-1 38

איור 2 : שיערוך פילוג זמן ההמתנה לקוחות 8-5 39

איור 3 : שיערוך זמן ההמתנה לקוחות 4-1 40

איור 4 : שיערוך זמן ההמתנה לקוחות 8-5 41

איור 5 : שיערוך זמן ההמתנה לקוחות 4-1 43

איור 6 : שיערוך זמן ההמתנה לקוחות 8-5 44

איור 7 : שיערוך זמן ההמתנה לקוחות 11-9 45

איור 8 : שיערוך פילוג זמן ההמתנה 46

איור 9 : משערך קפלן-מאייר לפילוג זמן ההמתנה לקוחות 4-1 49

איור 10 : משערך קפלן-מאייר לפילוג זמן ההמתנה לקוחות 8-5 50

איור 11 : משערך קפלן-מאייר לפילוג זמן ההמתנה לקוחות 12-9 51

איור 12 : משערך קפלן-מאייר לפילוג זמן ההמתנה לקוחות 16-13 52

איור 13 : משערך קפלן-מאייר לפילוג זמן ההמתנה לקוחות 18-17 53

איור 14 : משערך קפלן-מאייר לזמן ההמתנה לקוחות 4-1 54

איור 15 : משערך קפלן-מאייר לזמן ההמתנה לקוחות 8-5 55

איור 16 : משערך קפלן-מאייר לזמן ההמתנה לקוחות 12-9 56

איור 17 : משערך קפלן-מאייר לזמן ההמתנה לקוחות 16-13 57

איור 18 : משערך קפלן-מאייר לזמן ההמתנה לקוחות 18-17 58

איור 19 : משערך Censored MLE לזמן ההמתנה לקוחות 4-1 59

איור 20 : משערך Censored MLE לזמן ההמתנה לקוחות 8-5 60

איור 21 : משערך Censored MLE לזמן ההמתנה לקוחות 12-9 61

איור 22 : משערך Censored MLE לזמן ההמתנה לקוחות 14-13 62

איור 23 : פונקצית הסבלנות התיאורטית והמעשית 62

תקציר

מבוא

בעבודה זו אנו מתייחסים לפעולת תור בלתי נראה כתלות בסבלנות של הלקוחות. המוטיבציה לעבודה זו נובעת מהצורך האמיתי להבין את הציפיות איתן מגיעים לקוחות לשירותים המאופיינים בתורים בלתי נראים, על מנת שנוכל להגביר את שביעות רצונם מהשירות. הדוגמה השכיחה ביותר לשירותים אלו הינה מרכזי השירות (call-centers), החודרים כיום לעסקים רבים. לפי Direct Marketing Association השיווק והמכירות באמצעות מרכזי השירות הגיעו לסכום של $308 מיליארד או 44.6% מסך מכירות העסק לעסק (business-to-business) בשנת 1999 (סך כל המכירות באמצעות מרכזי השירות הגיע ל$538.3 מיליארד). בשנים האחרונות התרחבות שוק מרכזי השירות גדלה עוד יותר בעיקר בזכות חדירת האינטרנט. הרשת הינה בפירוש חזית חדשה עבור מרכזי השירות, אשר יכולים להגביר את האמינות והביטחון שבחווית הגלישה. באמצעות החיבור בין מרכזי השירות לרשת הוירטואלית, מגבירים את שביעות הרצון של הלקוחות מהשירות, ובסבירות טובה גורמים להם לחזור.

הקשר בין שביעות הרצון של הלקוחות מהשירות והצלחת השירות אינו דבר חדש, אך האם המעבר בין "זמן שווה כסף" להפחתת זמן ההמתנה בתורים גם הוא כה מיידי? לקוחות כבני אדם נוטים להסתגל לרמת השירות אותה הם מכירים. לדוגמה לקוח הרגיל להמתין דקה בתור, קרוב לוודאי ישקלל חווית המתנה של דקה וחצי בצורה שלילית. לעומתו, לקוח הרגיל להמתין שתי דקות יצא שבע רצון מהשירות. טענתנו המרכזית הינה, אם כן, כי שביעות הרצון מהשירות נקבעת לפי ההתאמה בין הציפיות איתן מגיע הלקוח לשירות, ובין רמת הביצועים של המערכת. ככל שנלמד להבין כיצד קובעים הלקוחות את הציפיות, כך נוכל לבצע החלטות מושכלות יותר בתהליך האופטימיזציה בין כמות השרתים במרכז השירות ובין הרווח המופק ממנו.

הנחות

לקוח הממתין בתור בלתי נראה מודע רק למצבו הוא בתור (כלומר האם קיבל שירות או לא) ופרק זמן ההמתנה בתור. אנו מאמינים כי הציפיות של לקוח משירות מסתמכות על האופן בו הוא תופס את ניסיונו הקודם במערכת. לכן, כאשר אנו מתייחסים למערכת של תורים בלתי נראים, עלינו להתייחס למידע המוגבל בו מחזיק הלקוח לגבי המערכת. כמו כן אנו טוענים שלקוח משתמש בכלל דגימות הזמן, המתארות את התנסויותיו הקודמות, על מנת לשערך את זמן ההמתנה הממוצע לו הוא מצפה. הנחות אלו הובילו אותנו לבחור בפרמטר זמן ההמתנה הממוצע כפרמטר המרכזי המשפיע על הציפיות וההתנהגות של לקוחות הממתינים בתור בלתי נראה.

מידול הבעיה

המודל שלנו מורכב משתי נקודות מבט, זאת של מנתח המערכת וזאת של המשתמש הפשוט. מנתח המערכת מנסה להבין את האופן בו מגיעים לקוחות להחלטה האם להמשיך להמתין או מתי לנטוש את התור, באמצעות הידע על פילוג זמן ההמתנה האמיתי בתור. לעומתו, הלקוח הפשוט מחזיק במידע חלקי על זמן ההמתנה – ניסיונותיו הקודמים במערכת. אנו מציעים שתי גישות, נפרדות אך קשורות, לתיאור התנהגות הלקוחות. גישה אחת מתייחסת לפונקצית סבלנות כוללת עבור תהליך ההחלטה שמבצעים כלל הלקוחות. גישה שניה מתייחסת לתהליך החלטה אינדיבידואלי שמבצע כל לקוח. שקלול ההחלטות האינדיבידואליות מוביל לקבלת פונקצית הסבלנות הכוללת. בשתי הגישות קיים פרמטר אחד, אשר מתאר את כל הידע בו מחזיקים הלקוחות – זמן ההמתנה הממוצע במערכת.

אנו מניחים לאורך העבודה כי המערכת הינה מערכת תורים מסוג M/M/m עם עזיבות, כאשר אורך התור אינו ידוע ללקוחות הממתינים ("תור בלתי נראה"). פונקצית הסבלנות אותה הזכרנו מוגדרת כהסתברות של לקוח הנבחר באקראי להישאר בתור t שניות לאחר הגעתו לתור. אנו מניחים כי פונקצית הסבלנות קשורה לאופן בו נתפסים ביצועי המערכת ע"י לקוחותיה, כלומר סבלנותו של לקוח המצפה מניסיונו כי המערכת תספק שרות מהיר ויעיל שונה מסבלנותו של לקוח שהתנסה בהמתנות ארוכות במערכת זו בעבר.

נחקרו שתי משפחות פונקציות המתארות את סבלנות הלקוחות: משפחת הפונקציות היורדות ב- x, ומשפחת הפונקציות העולות ב- x באופן חסום לינארית.

בנוסף, מוצע מודל החלטה רציונלי המתאר את תהליך ההחלטה המבוצע ע״י כל אחד מהלקוחות המתינים. אנו מניחים כי הלקוחות מאמינים בזמן המתנה מפולג אקספוננציאלית ומושפעים מפונקצית מחיר המתנה קמורה עולה. ההתנהגות התצריפית (aggregate) של הלקוחות הרציונלים מקושרת למשפחות הפונקציות המתארות את סבלנות הלקוחות.

הקשר ההדדי בין ביצועי המערכת ובין פונקצית הסבלנות יוצר משחק (במובן התיאורטי של ״תורת המשחקים״). קיום ויחידות של שיווי משקל במשחק מהווים חלק מרכזי בעבודה המובאת.

שיווי משקל

בבואנו לנתח את נקודות שיווי המשקל, עלינו להקדים ולהגדיר שיווי משקל זה. כפי שטענו, הלקוחות מחשבים את ממוצע זמן ההמתנה לפי המידע החלקי הנמצא בידיהם. כלומר באופן כללי פרמטר זה יכול להיות שונה מהערך האמיתי המאפיין את המערכת. אנו מגדירים את נקודת שיווי המשקל בתור הנקודה בה ערכי ממוצע זמן ההמתנה בהם מחזיקים כל הלקוחות וממוצע זמן ההמתנה האמיתי במערכת מתלכדים לערך זהה. חשוב לציין כי הפרמטר המדויק אליו אנו מתייחסים בתור ממוצע זמן ההמתנה הינו למעשה זמן ההמתנה הממוצע בהינתן שאכן נדרשת המתנה. כלומר, מצב בו לקוח מגיע למערכת ומיד מקבל שירות אינו נכלל בשקלול זמן ההמתנה הממוצע. הנחה זו מאד טבעית, מכיוון שאין משמעות אמיתית לפונקצית סבלנות במצב בו אין כל צורך בסבלנות.

שיווי המשקל כפי שהגדרנו אותו הוא:

$$E_x\left[v \middle| v > 0\right] = x$$

בצד שמאל של המשוואה מופיע ממוצע זמן ההמתנה המחושב לפי פונקצית פילוג זמן ההמתנה במערכת, ומצד ימין הפרמטר בו מחזיקים כל הלקוחות.

אנו מראים עבור משפחת פונקציות הסבלנות היורדות ב- x כי קיים שיווי משקל למערכת וכי שיווי משקל זה הינו יחיד. תוצאה זו מורחבת למקרה של פונקציות סבלנות עולות ב- x באופן חסום לינארית. בפרט, אנו משיגים קיום ויחידות של שיווי משקל עבור מודל ההחלטה ההגיוני, ע״י מציאת התוצר התצריפי של כלל ההחלטה האינדיבידואלי במשפחות הפונקציות המתוארות.

משערכים קטומים

עד כה חזרנו וטענו כי הלקוחות מסתמכים על המידע הבסיסי בו הם מחזיקים על מנת לקבוע את הציפיות שלהם או בעקיפין את הנכונות שלהם להמתין. אך עדיין לא התייחסנו לאופן בו משוקלל מידע זה לקבלת ממוצע זמן ההמתנה המשוערך. בפסקה הבאה נתייחס בצורה מפורטת יותר למידול תהליך הלמידה, אך קודם כל הבה נתרכז בשני משערכים אפשריים בהם בחרנו להשתמש: משערך קפלן-מאייר ומשערך נראות מירבית קטום (Censored MLE).

כאשר אנו רוצים לשערך את זמן ההמתנה הממוצע בתורים בלתי נראים עם עזיבות, עלינו להתייחס למידע בו מחזיקים הלקוחות במערכת תורים בלתי נראים. מידע זה הוא סדרת דגימות של זמני המתחלקת לשניים: דגימות של פרק הזמן עד הכניסה לשירות, ודגימות של פרקי זמן המתנה שהסתיימו בעקבות נטישת התור. הסוג השני מורכב למעשה מדגימות קטומות. דגימות אלו מדווחות על חסם תחתון לזמן ההמתנה הממוצע, ולא על פרק הזמן עצמו.

ברור כי איננו יכולים להתייחס לדגימות הקטומות באותה צורה בה אנו מתייחסים לדגימות הרגילות. כמו כן איננו יכולים להתעלם מדגימות אלו. נטישות מתרחשות כאשר פרקי זמן ההמתנה ארוכים מדי ביחס לסבלנותם של הלקוחות. על כן, התעלמות מהשוני של הדגימות הקטומות או ביטולן המוחלט יוביל להיסט של הממוצע המשוערך לערכים נמוכים יותר.

משערך קפלן-מאייר שייך לקבוצת המשערכים המתייחסים לדגימות קטומות. באמצעותו ניתן לשערך את פילוג זמן ההמתנה ומתוכו את ממוצע זמן ההמתנה. שיערוך פילוג זמן ההמתנה מתבצע לפי הנוסחה הבאה:

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{n_j - h_j}{n_j}$$

כאשר n_j הוא מספר הדגימות אשר לא הסתיימו ולא צונזרו עד לזמן t_j, ו h_j הוא מספר הדגימות שהסתיימו בזמן t_j.

יתרונו של משערך קפלן-מאייר בכך שאינו מניח דבר על פונקצית הפילוג, אלא משערך אותה, אך הסבירות שלקוחות יבצעו שיערוך בצורה זאת איננה גבוהה. משערך הנראות המירבית הקטומה, אומנם מניח הנחה שגויה שפונקצית הפילוג הינה אקספוננציאלית, אך קיימת סבירות גבוהה שלקוחות יניחו את אותה הנחה. חישוב משערך זה מתבצע בצורה הבאה:

$$\hat{\mu} = \frac{\sum_{i=1}^{m} x_i + \sum_{j=1}^{n-m} T_j}{m}$$

כאשר x_i היא דגימת פרק זמן המתנה עד קבלת שירות, ו T_i דגימת זמן נטישה.

מודל לתהליך הלמידה

לסיום אנו מציעים מודל לתהליך הלמידה הדינמי של הלקוחות. מודל זה מתבסס כמובן על ההנחה המרכזית שלנו – זמן ההמתנה הממוצע המשוערך משפיע השפעה חזקה על סבלנותם של הלקוחות. לפי המודל המוצע, לקוחות לומדים את התור באמצעות סדרת הדגימות בה הם מחזיקים. סדרת הדגימות מאפשרת ללקוח לשערך את זמן ההמתנה הממוצע, ולהחליט האם ומתי לנטוש את התור. כל החלטה יוצרת דגימה נוספת המצטרפת לסדרה הקיימת, ומעדכנת את הממוצע המשוערך.

המודל נבדק באמצעות סימולציה של מערכת תור M/M/1 עם עזיבות. לקוחות נכנסים למערכת וממתינים עד שאחד משני תנאים מתקיים: (1) הם מתקבלים לשירות, או (2) הם מאבדים את סבלנותם ומחליטים לעזוב. הלקוחות משערכים את זמן ההמתנה הממוצע באמצעות אחד משני משערכים, קפלן-מאייר או משערך נראות מירבית קטום. חוק העזיבה נקבע להיות פונקציה לינארית של ממוצע זמן ההמתנה המשוערך יחד עם אלמנט אקראי אדיטיבי.

תוצאות הסימולציות הראו התכנסות לשיווי המשקל התיאורטי עבור שימוש במשערך קפלן-מאייר. סימולציות עבור משערך MLE הקטום הראו היסט קבוע המתאים לחישוב האנליטי המתייחס להנחה המוטעית של פילוג זמן המתנה אקספוננציאלי.